The Open University

M140

Introducing statistics

Computer Book

The Open University

M140

Introducing statistics

Computer Book

This publication forms part of the Open University module M140 *Introducing statistics*. Details of this and other Open University modules can be obtained from Student Recruitment, The Open University, PO Box 197, Milton Keynes MK7 6BJ, United Kingdom (tel. +44 (0)300 303 5303; email general-enquiries@open.ac.uk).

Alternatively, you may visit the Open University website at www.open.ac.uk where you can learn more about the wide range of modules and packs offered at all levels by The Open University.

To purchase a selection of Open University materials visit www.ouw.co.uk, or contact Open University Worldwide, Walton Hall, Milton Keynes MK7 6AA, United Kingdom for a catalogue (tel. +44 (0)1908 274066; fax +44 (0)1908 858787; email ouw-customer-services@open.ac.uk).

The Open University, Walton Hall, Milton Keynes, MK7 6AA.

First published 2014. Second edition 2015.

Edited, designed and typeset by The Open University, using the Open University TEX System.

Printed in the United Kingdom by The Charlesworth Group, Wakefield.

SUP 04410 8

2.1

Contents

Introduction

This Computer Book describes all the computer activities you will be expected to complete during your study of M140. It contains detailed instructions in the use of Minitab, a statistical software package which you will be using to analyse data. It also contains activities that use interactive computer resources, which are designed to enhance your understanding of some of the statistical techniques introduced in M140.

Using this book

The work in the Computer Book assumes that you have reached various points in the M140 units. You should try to work through each subsection of the Computer Book at the point you are told to in the unit, although you may choose to delay this if you prefer. However, you should make sure that you complete these subsections before your study of the relevant unit is complete. You may not be able to complete TMAs and iCMAs until you do so.

About Minitab

Minitab is a statistical software package – that is, a software package designed to facilitate the statistical analysis of data. As you will discover as you study M140, many statistical procedures require calculations. In many cases, these calculations are not inherently difficult – but they can get very tedious, even with the help of a calculator! Statistical packages, such as Minitab, aim to eliminate the drudgery of statistics by carrying out such calculations quickly and easily.

About the interactive computer resources

The interactive computer resources are situated on the M140 website. These are designed to aid your understanding of particular parts of statistics covered in M140. They will allow you to explore properties of techniques without having to grapple with the mathematical details. Similarly to the activities involving Minitab, the activities involving interactive computer resources will not make sense until you reach the corresponding part in the unit. However, once you reach such a point in a unit, you are strongly recommended to engage with the corresponding interactive computer resource at that point, as it will assist you in your understanding of the rest of the unit.

A guided tour of Minitab

In this section, you will be taken on a quick tour of the Minitab environment.

You may prefer to watch the **Getting started with Minitab** screencast on the module website rather than working through Computer activity 1.

Computer activity 1 *Getting in and out of Minitab*

Run Minitab now: double-click on the **Minitab 17** icon on your desktop (or select 'Minitab 17 Statistical Software' from your list of programs).

An information panel telling you which version of Minitab is being used will flash up on the screen. (The '17' in the name 'Minitab 17 Statistical Software' indicates that it is version 17 of Minitab.) Then the opening screen will appear, as shown in Figure 1. Note that depending on the version of Windows on your computer, the screens that you see may differ slightly from those presented in this book.

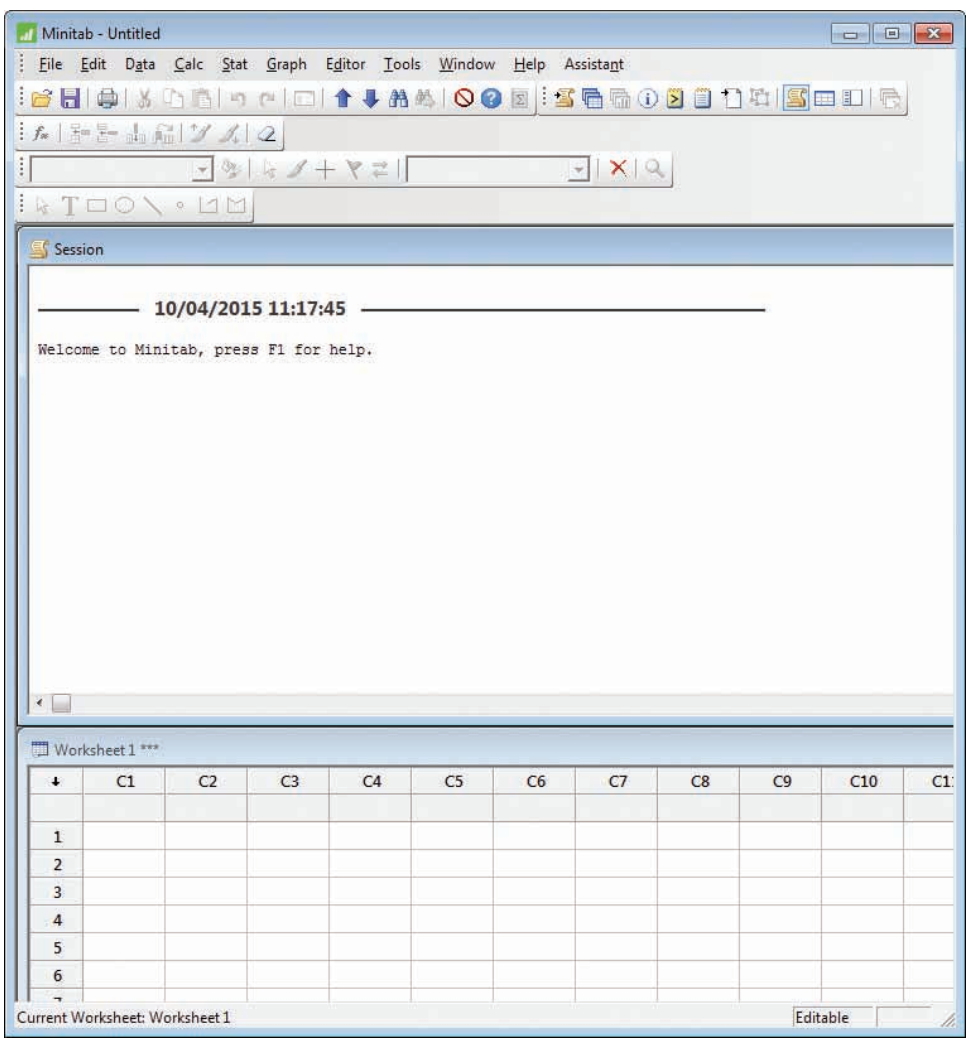


Figure 1 The opening screen in Minitab

The opening screen is similar to that of many Windows-based software packages: there is a menu bar at the top of the screen and a status bar at the bottom. Minitab opens with two windows between these: the top window is called the *Session* window and the bottom one is called the

Data window. The Session window is where the results are displayed, and the Data window is where your data are displayed.

Click on **File** in the menu bar to view the contents of the **File** menu. Notice that, as you move down the menu, when the pointer lingers on an item, a tool tip appears giving information about the item. An arrowhead pointing to the right on a menu item indicates the existence of a submenu. Spend a few minutes exploring the menus and their submenus. Do not be daunted by the large array of items listed in the menus and submenus. Only a small proportion of these will be used in M140.

(a) In which menu is **Calculator...** listed?

(b) In which menu is **Empirical CDF...** listed?

(c) In which menu is **1-Sample Z...** listed?

(d) Try to categorise the types of commands that are available in each menu.

To exit from Minitab, either click on the cross in the top right-hand corner of the opening screen, or click on **File** and then **Exit**. Do this now.

Computer activity 2 *Minitab worksheets*

Data in Minitab are stored in *worksheets*. When a worksheet is opened, the data it contains are displayed in a Data window.

The data described in Example 3 of Unit 1 (Subsection 2.2), on the numbers of large marine species, are contained in the worksheet named **bigfish.mtw**. Open this worksheet in Minitab now by doing the following.

The file name extension of a Minitab worksheet is **mtw**.

- Run Minitab.
- Choose **Open Worksheet...** from the **File** menu. This opens the **Open Worksheet** dialogue box.
- In the **Open Worksheet** dialogue box, navigate to the folder containing the M140 data files (that is, the files containing the datasets). The dialogue box should be similar to Figure 2.

Clicking on icons in the toolbar is a shortcut to opening some dialogue boxes. However, the icon for file opening is *not* the shortcut to the **Open Worksheet** dialogue box – it is the shortcut for a different dialogue box.

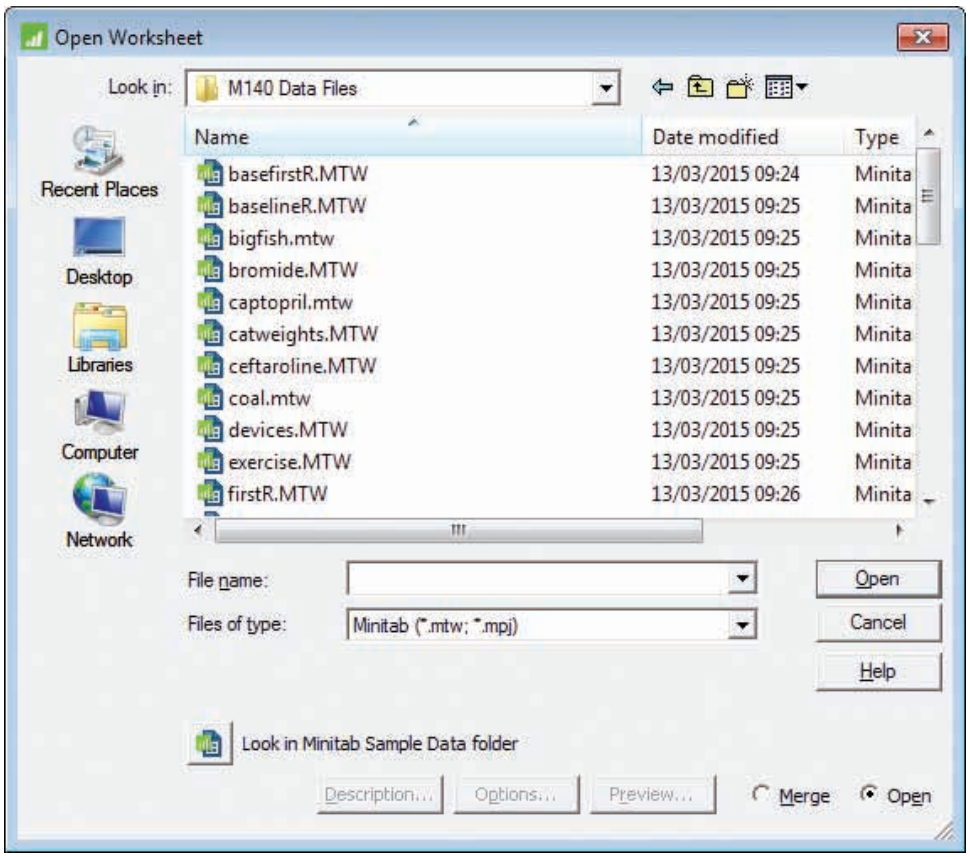


Figure 2 The **Open Worksheet** dialogue box

Minitab does not distinguish between upper- and lower-case letters in file names, so it does not matter which you use.

- Find **bigfish.mtw** and click on it to select the worksheet. If necessary, you can scroll to find the worksheet, or type **bigfish.mtw** in the **File name** field.
- Click on **Open** (or press `Enter` on the keyboard, or double-click on the file name).
- The message 'A copy of the content of this file will be added to the current project' will be displayed. Click on **OK**.

The data will now be displayed in a Data window such as that in Figure 3.

bigfish.mtw ***

	C1	C2	C3	C4	C
	year	new	total	model	
1	1829	1	101	101.104	
2	1830	3	104	103.281	
3	1831	1	105	105.420	
4	1832	1	106	107.521	
5	1833	2	108	109.584	
6	1834	1	109	111.610	
7	1835	2	111	113.600	
8	1836	0	111	115.554	
9	1837	4	115	117.473	
10	1838	2	117	119.358	
11	1839	9	126	121.210	
12	1840	1	127	123.028	
13	1841	1	128	124.814	
14	1842	1	129	126.568	
15	1843	0	129	128.290	
16	1844	5	134	129.982	

Figure 3 A Data window

Note first that the title of the Data window is **bigfish.mtw*****. The three asterisks indicate that this worksheet is the *active* one, which means it is the worksheet that Minitab will currently use. It is important to know which worksheet is the active worksheet if you have more than one worksheet open in Minitab at the same time.

Look at the data in the **bigfish.mtw** window. What do you think the rows represent? What do you think the columns represent?

The cells in a Data window contain values retrieved from a worksheet; you can also type values in directly. Note that a Data window only contains numerical values, not formulas.

So far, you have met two types of window in Minitab: the Session window and the Data window. There is another window that is present at all times, though it is minimised on the opening screen: the Project Manager window. You will be not expected to make use of this window in M140.

Computer activity 3 *Managing windows*

Within Minitab it is possible to have several windows open at one time. To demonstrate this, do the following.

- Open the worksheet **bigfish.mtw** in Minitab, if it is not already open.
- Next, open the worksheet **mpg.mtw** in Minitab. You should find **mpg.mtw** in the same folder as **bigfish.mtw**. (Look at the solution to this activity if you are unsure how to do this.)

  A second Data window opens, called **mpg.mtw*****. The three asterisks indicate that it has become the active window. This means that any commands you give will be applied to the data in **mpg.mtw**, not to **bigfish.mtw**.
- One way of making **bigfish.mtw** active once more is by clicking on it anywhere. Do this now.

The worksheet **mpg.mtw** contains the data on petrol consumption obtained after cleaning, given in Table 3 of Unit 1 (Subsection 3.3).

Sometimes the window you wish to make active is hidden beneath other windows, making it difficult or even impossible to click on. So another way of making a window active is by doing the following.

- Click on **Window** on the menu bar.
- At the bottom of the **Window** menu is a list of all the windows that are currently open in Minitab. (This includes the Session window and the Project Manager window.)
- On this list of windows, click on **mpg.mtw** to make it the active window once more.

Notice that making **mpg.mtw** the active window also brings it to the front. Using the **Window** menu is therefore a good way of finding windows that have become hidden behind other windows.

Spend a few moments navigating between windows, checking that you can make either worksheet the active one.

The guided tour of Minitab is now complete. Remember that you can exit from Minitab by selecting **Exit** from the **File** menu.

1 Scatterplots, stemplots and calculations

In this chapter, you will learn how to do some of the work in Unit 1 using Minitab. In Subsection 1.1, you will learn how to create scatterplots. Then in Subsection 1.2, you will learn how to perform numerical calculations on data using Minitab, and how to round your results to a specified number of decimal places. The production and interpretation of stemplots in Minitab is covered in Subsection 1.3. You will learn how to handle outliers and stretch stemplots by different amounts in Subsection 1.4. In Subsection 1.5, you will learn about histograms – another way of displaying data graphically, related to stemplots. Finally, in Subsection 1.6 you will learn how to print, paste and save your work.

This chapter is quite long and may take more than one study session to complete. However, you should aim to complete Subsection 1.2 and Subsection 1.6 each within a single session; later activities in these subsections depend on the state of Minitab resulting from earlier activities in the same subsection.

1.1 Obtaining a scatterplot

In this subsection, you will produce a basic scatterplot using Minitab. You will need the worksheets **bigfish.mtw** and **mpg.mtw**, so open them now in Minitab. The worksheet **bigfish.mtw** contains the data described in Example 3 of Unit 1 (Subsection 2.2) on the numbers of large marine species. The worksheet **mpg.mtw** contains the data on petrol consumption obtained after cleaning, given in Table 3 of Unit 1 (Subsection 3.3).

Look back at Computer activity 2 if you are unsure how to open worksheets.

Computer activity 4 *Plotting the discovery curve for large marine species*

In this activity, you will obtain a plot similar to the one in Figure 7 of Unit 1 (Subsection 2.2); that is, a plot of the total number of large marine species discovered by year. Begin by making the Data window for **bigfish.mtw** the active window.

Look back at Computer activity 3 if you are unsure how to make the **bigfish.mtw** worksheet the active window.

Scatterplots are produced using **Scatterplot...** from the **Graph** menu.

- Click on **Graph** and select **Scatterplot...** by clicking on it.
- A dialogue box will appear with several types of scatterplot. Click on the one called **Simple** to highlight it and then click on **OK** to select it.

Graph is found on the menu bar. As most procedures used in M140 start by selecting a menu from the menu bar, from now on 'from the menu bar' will be left implicit.

- The **Scatterplot: Simple** dialogue box will appear, as shown in Figure 4.

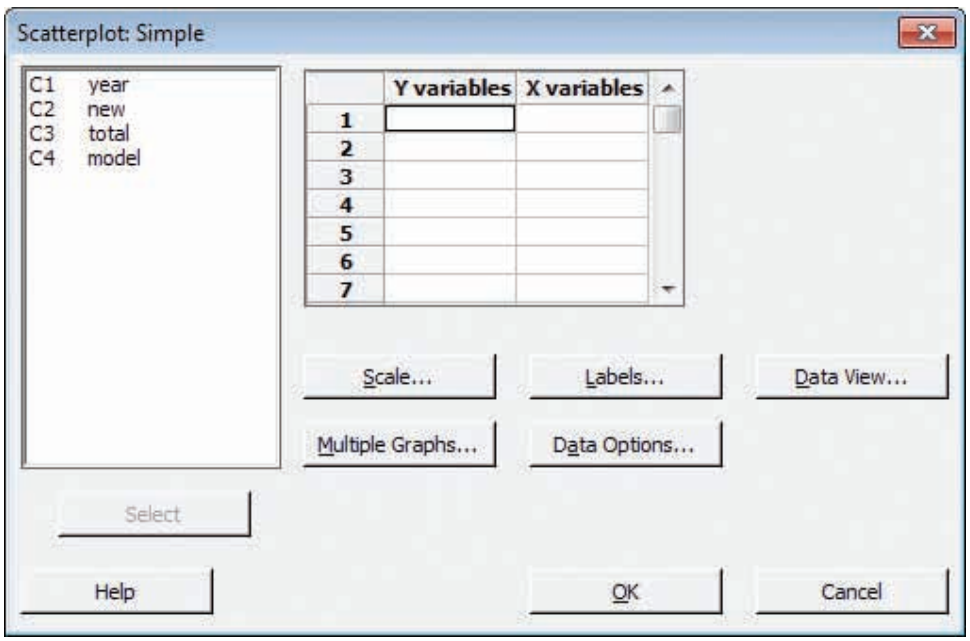


Figure 4 The **Scatterplot: Simple** dialogue box

On the left of the dialogue box are listed the variables in the worksheet **bigfish.mtw**. The variable `total` contains the total number of large marine species known about in each year. So, we want a plot of `total` against `year`, with `total` vertically and `year` horizontally. This corresponds to specifying that `total` is a Y variable and `year` is an X variable.

Which way round to specify the variables in a scatterplot is covered in Unit 5.

- To copy the variable names into the appropriate positions, start by selecting the cell in the first row of the **Y variables** field of the dialogue box by clicking on it (if it is not already selected).
- Click on the variable name `total` in the list in the dialogue box. The **Select** button will become active. Click on it, and the variable will be copied into the required cell in the **Y variables** field.

If by mistake you copy the variable into the wrong box, deselect it by clicking on the cell and pressing the delete button on your keyboard.

- The cell in the first row of the **X variables** field should have now become active. If not, click on it.
- Click on the variable name `year` in the list. Then click on **Select** to copy it into the cell in the **X variables** field.
- Finally, click on **OK**.

A Graph window will open, with the scatterplot in it as shown in Figure 5.

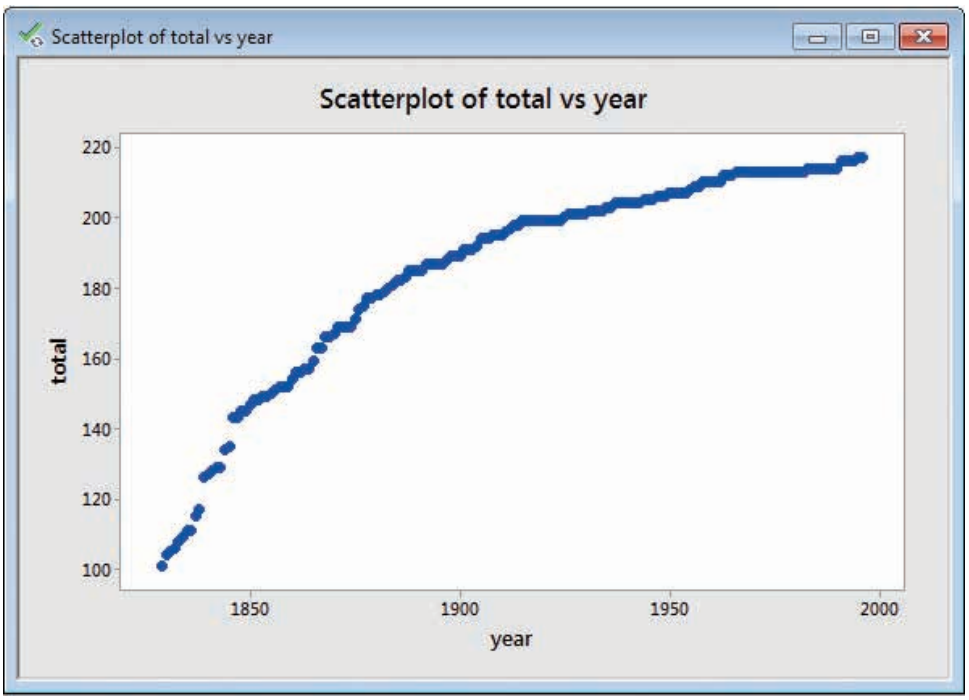


Figure 5 The discovery curve produced by Minitab

Minitab gives the scatterplot the title 'Scatterplot of total vs year'. This, along with other features of the plot, can be changed if you wish, but explanation of this will be deferred until Subsection 3.3 of this Computer Book (where we look at other plots).

Clicking on the **Window** menu will show that the Graph window is listed there. When you have a lot of windows open it's sometimes difficult to find the one you want as it might be hiding behind another; remember that clicking on its name in the **Window** menu will move it to the front.

The next activity will give you some more practice at selecting variables for analysis (the method is the same for all types of analysis) and some more practice obtaining scatterplots, using the petrol consumption data in **mpg.mtw**.

Computer activity 5 *Change in petrol consumption*

Start by making the Data window **mpg.mtw** active. The variable `mpg` gives the petrol consumption in miles per gallon at each of 34 periods between stops. It is of interest to see if the petrol consumption has changed in any way over time, represented here by the period number listed in variable `period`. One way to do this is to obtain a scatterplot of `mpg` against `period`.

(a) Obtain a scatterplot with `mpg` vertically and `period` horizontally. Note that when doing this you may find entries left in a dialogue box from the previous activity. Do not worry about this – when you enter a variable into a cell in the **Y variables** field or the **X variables** field, it will automatically replace any variable that is already there.

(b) Briefly interpret the scatterplot: over time, has the petrol consumption (measured in miles per gallon) tended to increase, decrease or stay the same?

In Minitab, windows containing graphs can be closed down without closing Minitab completely. Note that when you do this, you will be asked whether you want to save the graph in a separate file first (which you can choose to do if you wish).

1.2 Doing calculations in Minitab

In this subsection, we will use the petrol consumption data discussed in Section 3 of Unit 1. Open the worksheet **petrol.mtw**. There are five variables, in columns `C1` to `C5`. The variable `mileage2` is the mileage recorded at the end of the corresponding period in variable `period`. Thus, for example, the mileage recorded at the end of the first period was 112 616 miles. The variable `mileage1` gives the mileage recorded at the beginning of the period (so 112 616 is recorded for period 2). The variable `petrol` is the petrol (in litres) put in at the end of the corresponding period, and the variable `cost` is the cost (in pounds) of the petrol put in.

Computer activity 6 *Calculating the distance*

The difference between `mileage1` and `mileage2` is the distance travelled during a period.

Calculate this distance now using Minitab by doing the following.

- Calculations are done using **Calculator...** from the **Calc** menu. So, click on **Calc** and select **Calculator...**. The **Calculator** dialogue box will appear, with the variable names listed on the left.
- In the **Calculator** dialogue box, click in the **Store result in variable** field and type in `distance` using the keyboard. This instructs Minitab that the new variable is going to be called `distance`.
- Again in the **Calculator** dialogue box, click in the field called **Expression** and type the equation used to calculate the distance, namely `mileage2 - mileage1`. Note that extra spaces are ignored.

  (Alternatively, you could enter `C3-C2` or `'mileage2' - 'mileage1'` in the **Expression** field or, instead of typing the variable names, you could enter the expression by doing the following: double-click on `C3 mileage2` in the list of variable names, click on the '-' symbol in the keypad displayed in the dialogue box, and then double-click on `C2 mileage1`.)

The dialogue box should now look as in Figure 6.

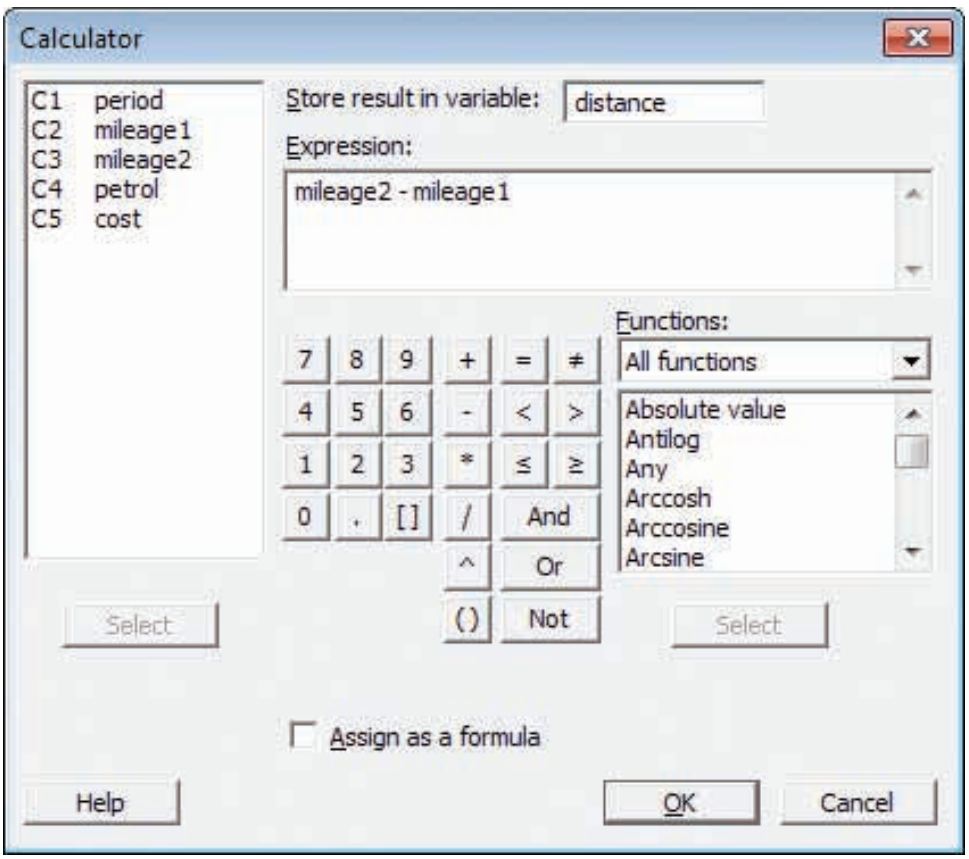


Figure 6 The **Calculator** dialogue box

- Now click on **OK**.

The new variable `distance` will appear in the worksheet, identified in Minitab by the next available column number, `C6`. Note that as the worksheet has been changed, when you come to exit Minitab you will be asked if you want to save the worksheet. For the moment, if do you need to exit Minitab, it is recommended that you click on **No**. However, you will need the modified worksheet from this activity in the next activity. So, if possible, carry straight on to the next activity without closing Minitab.

In Computer activity 6 you used Minitab to carry out a subtraction. Using the **Calculator** dialogue box it is possible to do many other types of calculation in Minitab. For example, calculations which include multiplication, division and taking powers, can be specified by using the symbols `*`, `/` and `^`, respectively. Also, you can enter expressions into the **Expression** field by using the keypad in the **Calculator** dialogue box instead of typing them as you did in Computer activity 6.

Practise doing a calculation in Minitab by completing Computer activity 7.

Computer activity 7 *Calculating the mpg*

From Subsection 3.3 of Unit 1 recall that petrol consumption can be measured as miles per gallon. Also,

$$\text{Petrol consumption (miles per gallon)} = \frac{\text{Distance travelled (miles)}}{\text{Petrol used (gallons)}}$$

and

$$\text{gallons} = \text{litres}/4.546\,09.$$

Together, this means that

$$\text{miles per gallon} = \frac{\text{Distance travelled (miles)} \times 4.546\,09}{\text{Petrol used (litres)}}.$$

It is assumed that your version of **petrol.mtw** contains the variable `distance` which you calculated in Computer activity 6. The worksheet **petrol2.mtw** is a copy of **petrol.mtw** modified in this way.

Use this formula to calculate a new variable in Minitab called `mpg` in your **petrol.mtw** worksheet. Note that you may need to remove entries left in a dialogue box from the previous activity: to do this, highlight them and press the delete button on your keyboard.

You will need the modified worksheet from this activity in the next activity, so if possible keep Minitab open and carry straight on to Computer activity 8.

In Section 3 of Unit 1, much emphasis was placed on using the correct rounding of numerical results in final answers (while keeping full accuracy for intermediate calculations, to avoid rounding errors). Minitab has a function that does the rounding for you. This is described in the next activity.

Computer activity 8 *Rounding in Minitab*

The worksheet **petrol3.mtw** is a copy of **petrol.mtw** with the extra variables `distance` and `mpg`.

The variable `mpg` in the **petrol.mtw** worksheet contains the fuel consumption in miles per gallon, calculated to more than ten decimal places and displayed to four decimal places. In Example 9 of Unit 1 (Subsection 3.3), it was concluded that the correct accuracy should be three significant figures. Minitab only rounds using decimal places, and not significant figures. So to round to a certain number of significant figures you need to work out how many decimal places this would be. In this case we need 26.8844 to be rounded to 26.9, so we need to round to one decimal place. Do this now by doing the following.

- Select **Calculator...** from the **Calc** menu to open the **Calculator** dialogue box. (A shorthand way of writing this is **Calc** > **Calculator**.)
- Call the rounded variable `roundedmpg` by entering this variable name in the **Store results in variable** field.

- Under the **Functions** field within the **Calculator** dialogue box, there is a long list of functions. Find the function called `Round`. You can scroll up and down the long list until you find it. You can also shorten the list by opening the drop-down menu beside `All functions` and selecting `Arithmetic`. This results in only 'Arithmetic' functions, of which `Round` is one, being displayed in the list.

  Once you have found `Round`, click on it.
- The **Select** button just beneath the **Functions** field will become active, and the exact form of the `Round` command will be displayed as `ROUND(number,decimals)`. The numbers we want to round are in column `mpg`, and we want to round them to one decimal place. Click on **Select** to put the command into the **Expression** field. (Again, you may need to delete entries in the dialogue box from the previous activity.) Replace `number` by `mpg` and `decimals` by `1`. The dialogue box should be as shown in Figure 7.

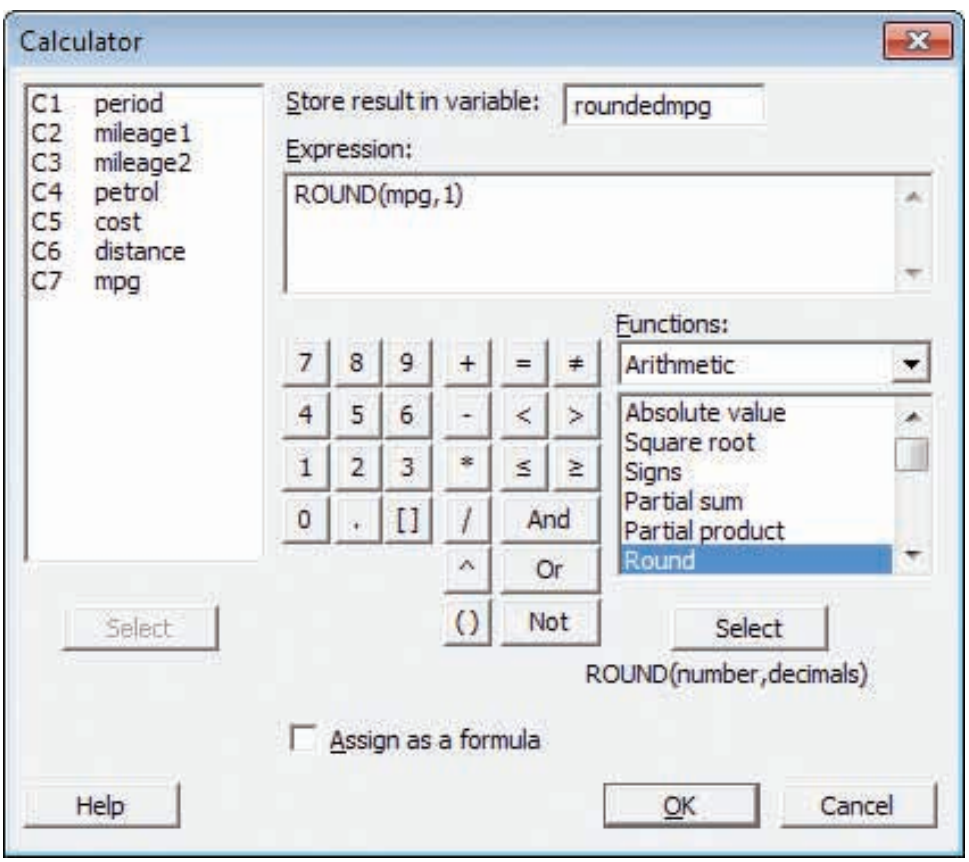


Figure 7 Rounding with Minitab

- Now click on **OK**.

The rounded values will appear in a new column, `C8`, called `roundedmpg`. For example, the first entry in this column is 26.9. Note that Minitab uses the same rounding convention as was described in Subsection 3.2 of Unit 1.

You can combine functions and arithmetic expressions if required. The next activity will give you practice at doing this.

Computer activity 9 *The cost per mile*

One question of interest is the fuel cost of running the car, per mile travelled. This may be calculated in Minitab as `cost/distance`, in £ per mile. Now, `cost` has four significant figures, and `distance` has three, so the result should be rounded to three significant figures. For the first period, the running cost is $49.89/266 \simeq 0.1876$, so three significant figures corresponds to three decimal places. This can be done in one go in Minitab using the function `ROUND(cost/distance,3)`.

Use Minitab to calculate this quantity, to be called `costpm` (cost per mile).

Minitab has many other functions which you can use within the **Calculator** dialogue box. They can all be used in a similar way to the `ROUND` function.

1.3 Stemplots

In Minitab, stemplots are called *stem-and-leaf plots*. In this subsection you will learn how to obtain stemplots in Minitab.

Computer activity 10 *Obtaining a stemplot in Minitab*

These data were described in Activity 19(a) of Unit 1 (Subsection 5.2).

Run Minitab now and open the worksheet **coal.mtw**. This contains the data on coal production in the UK in 1970/71. There is a single variable, `production`, which is the coal production by region in thousand tonnes. Produce a stemplot of these data using Minitab by doing the following.

- Choose **Stem-and-Leaf...** from the **Graph** menu. The **Stem-and-Leaf** dialogue box will open.
- Enter the variable `production` into the **Graph variables** field (either by typing it in, clicking on the variable name then on `Select` or by double-clicking on the variable name in the field on the left-hand side). Leave all the other fields blank. The **Stem-and-Leaf** dialogue box should look like the one shown in Figure 8.

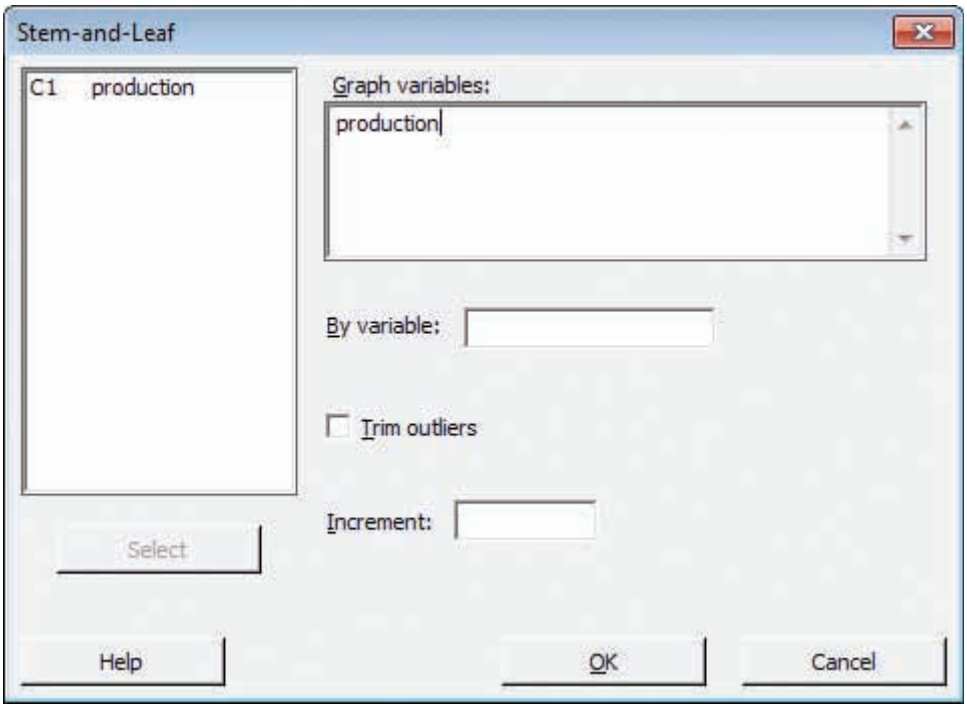


Figure 8 The **Stem-and-Leaf** dialogue box

- Click on **OK**.

The following will appear in the Session window. (You may need to scroll up to see the top.)

```
Stem-and-leaf of production  N  = 18
Leaf Unit = 100


 1    1    0
 1    2
 1    3
 2    4    7
 5    5    128
 8    6    128
(2)   7    57
 8    8    038
 5    9    447
 2    10   5
 1    11
 1    12   0
```

The stemplot is displayed in 12 rows and 3 columns. Each row is a level of the stemplot. The rightmost column gives the leaves, and the middle column is the stem. Together, these two columns form the stemplot: this is identical to that of Figure 31 in Unit 1.

At the top, the batch size is listed, along with the units the leaves are measured in (the `Leaf Unit`), expressed in terms of the original data. Here the batch size is 18 and the leaf unit is 100, so for example the leaf '7' on level 4 of the stemplot represents 4700 thousand tonnes.

The leftmost column is called the `Count` column. It represents the cumulative number of values, starting from the end that gives the lower cumulative number. So, for example, at level 4 the entry in the `Count` column indicates that there are 2 values at or below level 4 (one at level 1 and one at level 4). Similarly, at level 9, it tells you there are 5 values at level 9 and above (three at level 9, one at level 10 and one at level 12). The leaves at level 7 would need to be split between the two tails, so Minitab simply reports the number of leaves at that level, and encloses that number in brackets. This, incidentally, is the 'middle level' where the median is located. (Occasionally the median is the average of the last leaf on one level and the first leaf on the next level with a leaf. In such cases none of the numbers in the `Count` column will have brackets round them.)

The purpose of the `Count` column is to enable you to see quickly at what level (or between which levels) the median is: it's either where the number in brackets is (if there is one), or on one of the levels which have the highest `Count`.

Use this stemplot to find the median coal production in 1970/71.

By default, stemplots are placed in the same Session window. (This is unlike the scatterplots you have created, which each go in a separate window.) However, you might have to scroll up or down in the Session window to see the stemplot on the screen.

Let's have a look at another dataset.

Computer activity 11 *A stemplot of shot-put results*

These data were presented in Exercise 6 of Unit 1 (Exercises on Section 4).

Open the worksheet **shotput.mtw**. There are two variables, `distance` and `pool`; they represent the shot-put results for 15 senior male athletes in the UK, grouped in two pools. The variable `distance` contains the athletes' best throws, in metres.

If the **Graph variables** field is not initially empty, double-clicking on `distance` will replace any highlighted entry in this field.

(a) Obtain a stemplot of `distance`: on the **Graph** menu select **Stem-and-Leaf...**, enter `distance` in the **Graph variables** field, leave all other fields blank, and click on **OK**. What has now appeared in the Session window? Use the resulting output to calculate the median best throw by these athletes.

(b) Obtain the **Stem-and-Leaf** dialogue box again. Make the **By variable** field active (by clicking in it) and insert the variable `pool` in that field. Click on **OK**. What now appears in the Session window?

(c) Use the output you obtained in part (b) to calculate the median best throw in each of the two pools.

The next activity will give you some more practice at obtaining stemplots.

Computer activity 12 *A stemplot of race times*

The Minitab worksheet **race.mtw** contains data on the times, in seconds, taken in the finals of 400-metre races. There are two variables: `time` and `sex`. For the variable `sex`, 1 denotes women and 2 denotes men. Open the worksheet **race.mtw** now.

These data were discussed in Activity 19(b) of Unit 1 (Subsection 5.2).

(a) Obtain a stemplot of `time`, with women and men combined. What is the median?

(b) Now obtain separate stemplots for women and men. Use these stemplots to obtain the separate median times for women and for men.

1.4 Customising stemplots

In this subsection you will learn how to customise stemplots by adjusting the number of levels, and by listing the outliers separately if required.

Computer activity 13 *A stemplot of television prices*

Activity 17 of Unit 1 (Subsection 5.1) introduced some data on the price of 26 digital televisions in February 2012. These data are given in the Minitab worksheet **tvs.mtw**. Open this worksheet in Minitab now.

Produce a stemplot of these data in Minitab. What do you notice about the stemplot?

In Computer activity 13 you have seen that Minitab sometimes produces stretched stemplots. In Minitab it is also possible to control the *amount* of stretching that Minitab applies to a stemplot.

Computer activity 14 *Controlling the stretch on a stemplot*

In Computer activity 10 you produced a stemplot of coal production in 1970/71. That stemplot was not stretched automatically by Minitab. Suppose now that a new stemplot is required, one where the leaf unit is 1000 (measured in thousand tonnes) and each level is split into five parts.

- Open **coal.mtw** in Minitab, if it is not already open, and make sure it is the active window.
- Obtain the **Stem-and-Leaf** dialogue box (**Graph** > **Stem-and-Leaf**) and enter `production` in the **Graph variables** field.

In the **Stem-and-Leaf** dialogue box, notice that there is a field labelled **Increment**. By entering a number in this field it is possible to control the amount of stretching Minitab applies to a stemplot. The number given in this field is assumed to be in the same units as the data in the worksheet. So, in this case, the number will be in terms of thousand tonnes. Its value is given by the separation between two successive rows on the stem. In this case, we require leaf units of 1000 and each level split into five parts. This means that two different digits are possible on each part of each level (first part: 0 or 1, second part: 2 or 3, third part: 4 or 5, fourth part: 6 or 7, fifth part: 8 or 9). So there is $2 \times 1000 = 2000$ (thousand tonnes) separating successive rows on the stem.

- Enter `2000` in the **Increment** field.
- Click on **OK**.

The finished stemplot is as follows.

```
Stem-and-leaf of production  N   = 18
Leaf Unit = 1000


 1    0  1
 1    0
 5    0  4555
(5)   0  66677
 8    0  888999
 2    1  0
 1    1  2
```

Computer activity 15 *Trimming a stemplot*

(a) In Subsection 4.2 of Unit 1 you saw that a stemplot can be made more compact by listing outliers outside the main body of the stemplot. Create such a stemplot for the data on TV prices used in Computer activity 13 by doing the following.

- Open the file **tvs.mtw** in Minitab, if it is not already open, and make sure it is the active window.
- Bring up the **Stem-and-Leaf** dialogue box (**Graph** > **Stem-and-Leaf**).
- In the **Stem-and-Leaf** dialogue box, enter `price` in the **Graph variables** field and make sure that the **Increment** field is blank.
- Notice that in the **Stem-and-Leaf** dialogue box there is an option to **Trim outliers**. Select **Trim outliers**. (An option is selected when there is a tick beside it. It is possible to switch between selecting and not selecting an option by clicking on the box beside it.)
- Click on **OK**.

(b) Compare the stemplot you obtained in part (a) with the one you obtained in Computer activity 13 and that displayed in Activity 17 of Unit 1 (Subsection 5.1). Which observations has Minitab identified as outliers? Does this match the outliers picked out in Activity 17 of Unit 1 (Subsection 5.1)?

(c) Does the amount of squeezing/stretching chosen by Minitab in the stemplot in part (a) match that used in Computer activity 13?

In the next activity, you will get some practice with customising stemplots. Note that not all values are admissible as increments in the **Increment** field. This is because the maximum number of different types of leaf available at each level must be the same. For example, an increment that splits levels into two parts is admissible as the number of different possible leaves for each part is the same (corresponding to leaves that are either one of 0, 1, 2, 3, 4 or one of 5, 6, 7, 8, 9). In contrast, an increment that splits levels into three parts is not admissible as one part will have four different possible leaves but the other two parts only three different possible leaves. You can enter any value you want, but if it's not admissible then Minitab will automatically select the next highest admissible value. Also note that Minitab will not trim outliers when it produces separate stemplots for subgroups of the data.

One reasonable approach is just to enter a number and see what happens!

Computer activity 16 *Stretching, squeezing and trimming high salaries*

The Minitab worksheet **hipaid.mtw** contains data on the median salaries of the 20 highest-paid professions. Open this spreadsheet now and make sure it is the active window. There is one variable, `salary`, giving the median annual salary in £.

These data were introduced in Example 12 of Unit 1 (Subsection 4.2).

(a) Obtain the default stemplot for `salary` (i.e. leave the **Increment** field blank and **Trim outliers** unselected). How many outliers would you say there are? What is the leaf unit? How many parts have the levels been split into?

(b) Now obtain a second stemplot with the outliers trimmed, leaving everything else as before. What do you observe? How many outliers have been trimmed?

(c) Keeping the outliers trimmed, enter a value in the **Increment** field so that each of the levels only has one part (i.e. not split). What increment did you use? What values of the increment will produce the same result?

(d) Now adjust the increment so that each level is split into five parts (keeping the outliers trimmed). What increment did you use? What values of the increment will produce the same result?

1.5 Histograms

In this subsection, you will be shown a different way of presenting data graphically – the resulting diagrams are called *histograms*. Histograms do not play a major role in M140, but they are very common in some other contexts, and you will almost certainly have seen examples elsewhere. As you will see, histograms have something in common with stemplots.

With a stemplot you know that you can get an idea of which data values are relatively common by looking at the number of leaves that appear in the different rows – and you do not have to count them to get an impression, you can just look at the length of the rows. For instance, look again at this (stretched) stemplot of data on coal production for 18 regions, that you produced in Computer activity 14.

```
Stem-and-leaf of production  N   = 18
Leaf Unit = 1000


 1    0  1
 1    0
 5    0  4555
(5)   0  66677
 8    0  888999
 2    1  0
 1    1  2
```

If you just look at the length of the rows, and do not worry too much about the exact values of all the leaves, you can still see that the most common production figures are around 8000 or 9000 (thousand tonnes), and there are very few values below 4000.

You can think of a histogram as a diagram in which a stemplot is turned on its side, and the rows of the stemplot are replaced by bars – with the length of a bar representing the number of leaves in a row. Because the plot has been turned around, the vertical bars of the histogram correspond to the horizontal rows of the stemplot.

Figure 9 shows a histogram that corresponds to the stemplot above.

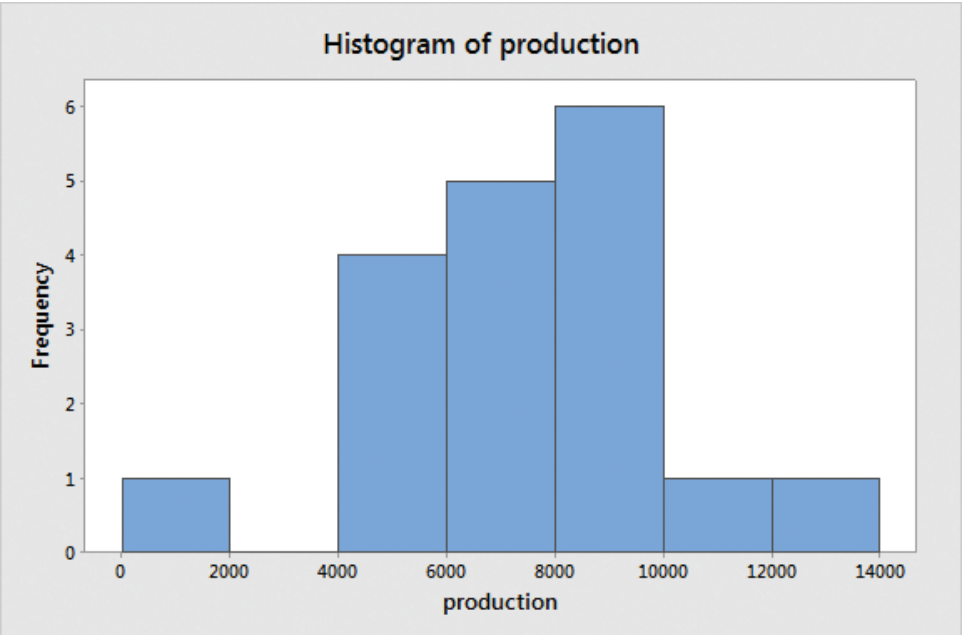


Figure 9 A histogram of coal production data, drawn by Minitab

There is a horizontal scale along the bottom showing the production figures, which can be interpreted in the same way as a scale on a scatterplot. The first bar on the left runs from 0 to 2000 on the horizontal scale, so it relates to regions where production was between 0 and 2000 (in thousand tonnes). The height of that bar is one unit on the vertical scale. That scale is labelled 'Frequency', which means that the bar represents the number of regions that have a production figure between 0 and 2000. For this bar, the frequency is 1 because there is just one such region. That matches up with the stemplot, where this one region is in the top row of the plot, with a leaf value of 1 corresponding to a production of 1000.

Computer activity 17 *Reading information from a histogram*

Look at the third bar from the right in Figure 9. What range of coal production figures does it represent? How many regions fall into that range? Does that match up to what you can see in the stemplot?

So, in this case at least, the histogram presents essentially the same information as the stemplot, but in a different way. Arguably, it is easier to see the shape of the distribution of the data in the histogram, because you are not distracted by the numbers that make up the leaves. But those numbers do provide extra information that you cannot get from the histogram. For instance, for the third row from the bottom of the stemplot, corresponding to the third bar from the right on the histogram,

the leaves are `888999`, showing that three of these six values are 8000 and the other three are 9000 (after rounding). All you can tell from the histogram is that they are all between 8000 and 10 000.

Histograms are not very difficult to draw using Minitab, though the results may not always be quite what you expect – as you will see in Computer activity 18.

Computer activity 18 *Drawing and customising a histogram using Minitab*

Produce a histogram of the coal production data in Minitab, by doing the following.

- Open **coal.mtw** in Minitab, if it is not already open, and make sure it is the active window.
- Click on **Graph** and then **Histogram. . . .**
- The **Histograms** dialogue box will appear. This has four small pictures, labelled **Simple**, **With Fit**, **With Groups**, and **With Fit and Groups**. We will only be using the **Simple** version in M140. Make sure that **Simple** is selected, and click on **OK**.
- The **Histogram: Simple** dialogue box appears. Copy the variable `production` into the **Graph variables** field. The completed dialogue box should be the same as Figure 10.

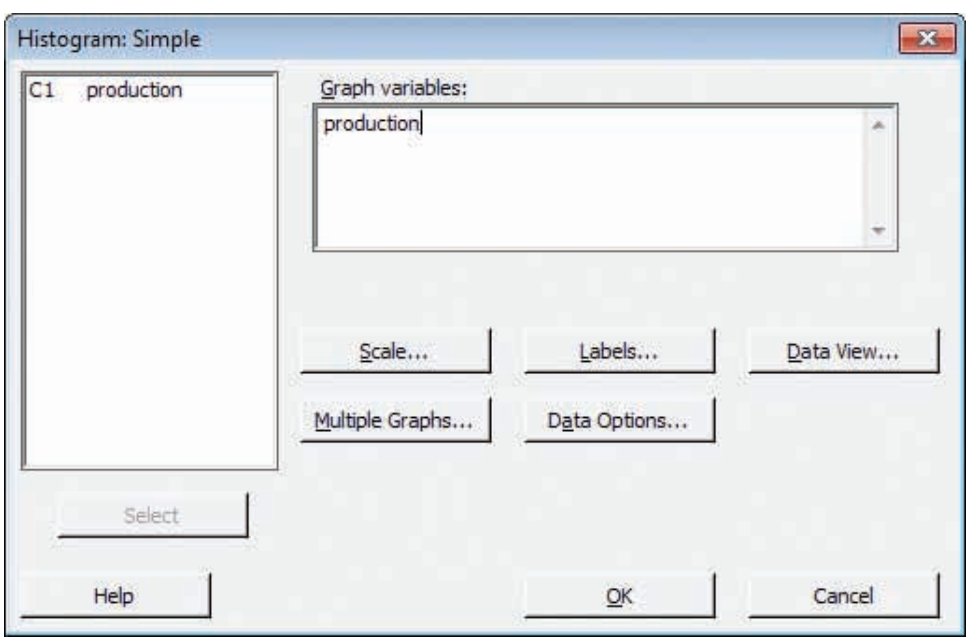


Figure 10 The **Histogram: Simple** dialogue box

- Click on **OK**.

A new window opens containing a histogram, as shown in Figure 11.

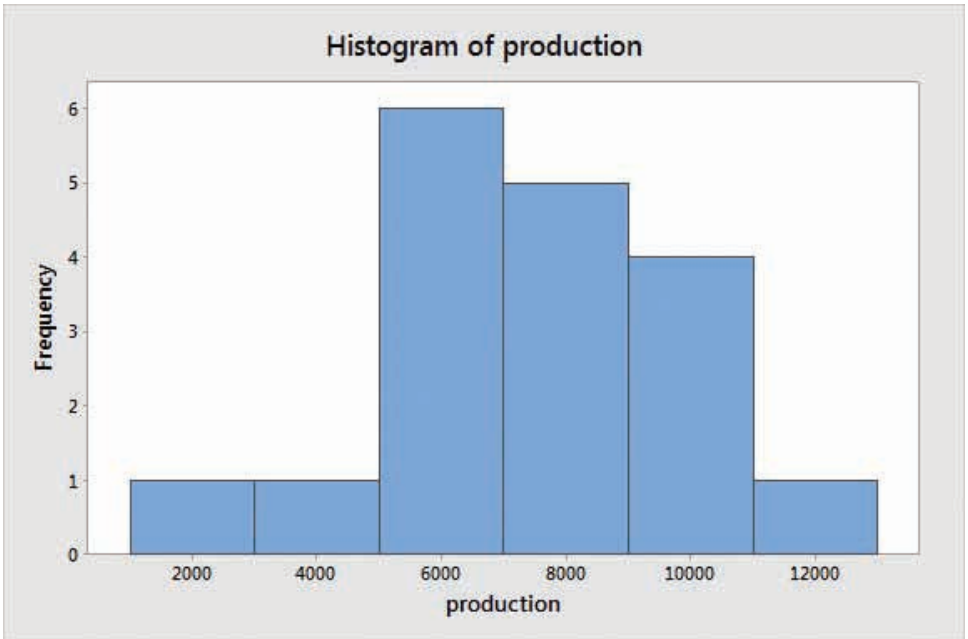


Figure 11 The default histogram of coal production, as produced by Minitab

You might have been surprised to notice that the histogram in Figure 11 is not the same as the one in Figure 9, although it is based on the same data. The reason is that the bars are differently positioned along the horizontal axis, and so cover different ranges of values of coal production. In Figure 9, the leftmost bar corresponded to production figures between 0 and 2000. In Figure 11, it is perhaps not so clear to see exactly which values this bar corresponds to – because the positions of the edges of the bar are not labelled on the scale. In fact, the bar is for production figures between 1000 and 3000. Again, the vertical scale says that there is only one such value. This does correspond to the stemplot: as before, the '1' leaf on the top row of the stemplot. But in this histogram, the individual bars do not correspond exactly to the rows in the stemplot, as was the case in Figure 9. That does not make the new histogram wrong – but it does make it different. In general terms, both histograms do give the same impression of the overall shape of this batch of data; however, the details are rather different.

If you want to change where Minitab puts the edges of the bars in a histogram, you *can* do so. First you must draw the default histogram (as in Figure 11). Then you can edit where the bars go, as follows.

- In the histogram of coal production, double-click on the horizontal scale – it is probably easiest to do this by double-clicking on one of the numerical labels for that scale. This brings up the **Edit Scale** dialogue box.

  Alternatively, click on **Editor** (in the menu bar at the top) and then **Select item**. This brings up a submenu of items on the histogram that can be selected for editing. Click on **X Scale**. Notice that marks are placed on the horizontal scale to show that it has been selected.

Now click on **Editor** again and click on **Edit X Scale...**. This brings up the **Edit Scale** dialogue box.

- In the **Edit Scale** dialogue box, click on the **Binning** tab (at the top) to open it. (The rather strange name is because the ranges defining the histogram bars are called 'bins'.)

  At the top of the resulting dialogue box, there is a field called **Interval Type** – and in that field, **Midpoint** is selected. That means that it is the middles of the histogram bars that are labelled on the histogram, as in Figure 11. The other alternative in this field is **Cutpoint**, which labels the edges of the bars instead, as in Figure 9. Click on **Cutpoint** to select it, and then on **OK** to make the change.

The histogram changes, to that in Figure 12.

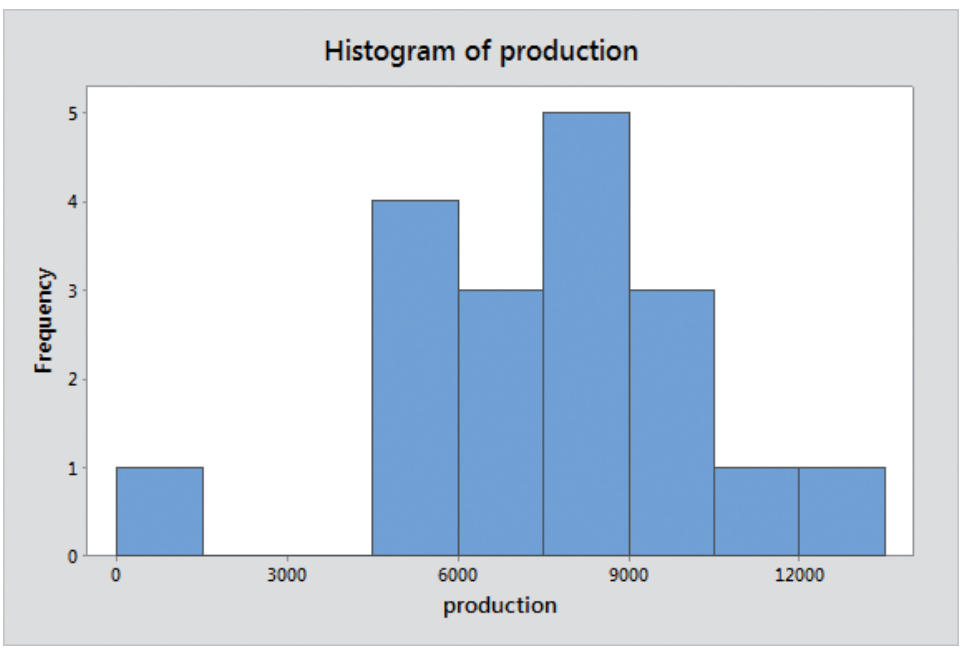


Figure 12 Histogram of coal production with **Interval Type** set to **Cutpoint**

The edges of the bars are labelled, or at least some of them are. But the plot still does not look the same as Figure 9. Minitab has chosen a different set of 'bins'. For example, the first bar on the left now runs from 0 to 1500 (thousand tonnes), whereas in Figure 9 the leftmost bar ran from 0 to 2000. Furthermore, the impression that Figure 12 gives of the shape of the batch of data is rather different to the impression given by either Figure 9 or Figure 11. There is a 'dip' in the main body of the data, at 6000 to 7500, that is quite noticeable in Figure 12 but did not show up in the other two histograms.

To get Figure 9 exactly, do the following.

- Obtain the **Edit Scale** dialogue box again, for the horizontal scale, either by double-clicking on the scale again, or by using the **Edit** menu as described above.

- In the **Edit Scale** dialogue box, open the **Binning** tab. Check that **Cutpoint** is still selected in the **Interval Type** field.
- Now you need to tell Minitab explicitly where you want the boundaries between the 'bins' to come, rather than letting the software choose them automatically. In the **Interval Definition** field, click on **Midpoint/Cutpoint positions** to select it. The field below contains the current positions of the boundaries, with spaces between them. Select all of these and delete them. In their place, type

  ```
  0 2000 4000 6000 8000 10000 12000 14000
  ```

  These are the right positions to match Figure 9. The completed dialogue box should look as in Figure 13.

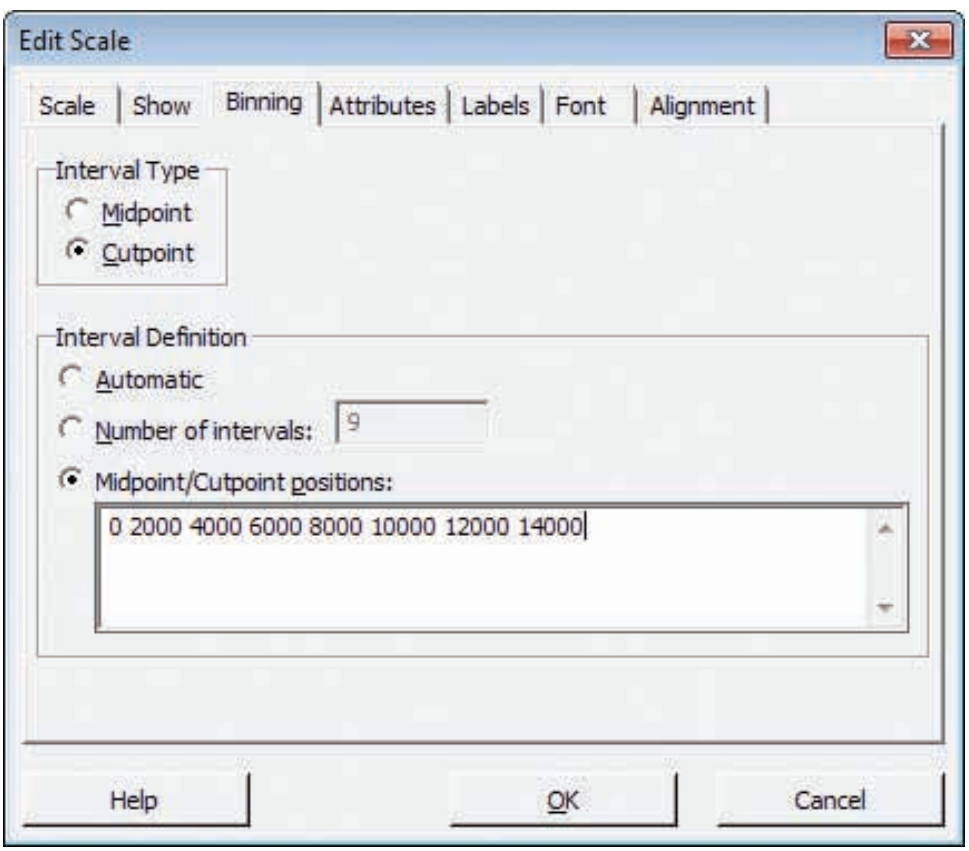


Figure 13 The **Edit Scale** dialogue box

- Click on **OK**.

The histogram changes again, and now it should look like Figure 9 at last.

You might be wondering where Minitab would put a coal production figure, in Figure 9, that was exactly 2000. There is no such figure in the data, but what if there were? Would it contribute to the bar for 0 to 2000, or the bar for 2000 to 4000? Either choice would be a reasonable one.

However, Minitab always chooses to put it in the bar that corresponds to the lower values on the horizontal scale, so in this case it would go in the 0 to 2000 bar.

Histograms have many uses for plotting data that could, in theory, take any value at all between a maximum and a minimum. (You should not use a histogram to plot, say, data on the number of children in families. Those data will consist of relatively small whole numbers, 0, 1, 2, and so on. Plotting them as a histogram, with bars that have no gaps between them, might give the impression that a family could have 2.5 or 3.2 children, which is not possible. There are similar plots, called 'bar charts', for such data.) In particular, they are helpful when there are too many values to fit comfortably on a stemplot.

However, there are choices to be made when drawing a histogram, particularly about where the 'bins' should begin and end. As you have seen, such choices can change the way a histogram looks – this is particularly true in small batches of data. In M140, you will not be expected to work out a good choice of these 'bins' yourself.

1.6 Printing, pasting and saving your work

So far in this chapter you have concentrated on using Minitab to analyse data. This section concentrates on the more administrative-type tasks in Minitab: printing, pasting and saving your work.

In Computer activity 19, you will create some work in Minitab to be printed, pasted and saved. Do not end your Minitab session after this activity; you will need the worksheet and output for the rest of the activities in this section.

Computer activity 19 *Exploring petrol consumption*

In Subsection 1.2, you used Minitab to calculate petrol consumption for the data given in Section 3 of Unit 1. These data, including the variables `distance` and `mpg`, are given in the worksheet **petrol3.mtw**. Open this worksheet in Minitab now.

(a) Obtain the default stemplot for `distance` – that is, the stemplot that is obtained when the **Trim outliers** option is not selected and the **By variables** and **Increment** fields are left blank. Briefly comment on the shape of the stemplot.

(b) Obtain a scatterplot for `mpg` against `distance`. As the distance between stops increases, does the petrol consumption (measured in miles per gallon) appear to go up, go down or stay about the same?

Computer activity 20 *Printing a Data window*

Computer activity 19 used the worksheet **petrol3.mtw**. A printout of this worksheet can be obtained by doing the following.

- Start by making sure that the **petrol3.mtw** window is active.
- Click **File** and select **Print Worksheet...**.
- The **Data Window Print Options** dialogue box provides you with various options for printing. If you are printing out a worksheet for inclusion in a (paper) TMA, it is recommended that you ensure that the options **Print Row Labels**, **Print Column Labels** and **Print Column Names** are selected.
- Click on **OK**. The **Print** dialogue box then opens, allowing you to choose your printer options. Check these are correct before clicking on **OK**.

Computer activity 21 *Printing output from the Session window*

The stemplot produced in Computer activity 19 is displayed by Minitab in its Session window. Output such as this can be printed. Print this stemplot by doing the following.

- First make the Session window active.
- Highlight the part of the output in the Session window that you wish to print – in this case the stemplot produced in Computer activity 19. (To highlight a section of text, you could drag your mouse over it.)
- Click **File** and select **Print Session Window...**.
- The **Print** dialogue box then opens, allowing you to choose your printer options. Make sure that the **Print range** is set to **Selection**. Check the other options are correct before clicking on **OK**.

Computer activity 22 *Printing a Graph window*

The scatterplot produced in Computer activity 19 is displayed by Minitab in a Graph window. Print this scatterplot by doing the following.

- First make the Graph window containing the scatterplot active. (If necessary use the **Window** menu item to find it.)
- Click **File** and select **Print Graph...**.
- The **Print** dialogue box then opens, allowing you to choose your printer options. Check these are correct before clicking on **OK**.

Although printing worksheets and output in this way is very simple, you may prefer to paste worksheets or output into a word-processor document and then print the document. This is described in the next three activities.

Computer activity 23 *Copying and pasting part of a Minitab worksheet*

You can copy cells from a Minitab worksheet and insert them into a word-processor document. Try this now for the data in **petrol3.mtw** by doing the following.

- Open a word-processor document and type in 'The data on petrol consumption is as follows.'.
- Switch to Minitab and in **petrol3.mtw** highlight all the columns with data in them, by clicking on the cell labelled '`C1`' and dragging across. This will ensure that all of the data will be copied into your word-processor document. (Note that you will not be able to highlight grey cells at the left-hand side of the worksheet.)

  If you only wish to copy some of the cells in the worksheet, just highlight the particular cells that you wish to copy.
- Select **Edit** and then **Copy Cells**. (Alternatively, press `Ctrl`+`C`.)
- Now switch to your word-processor document (perhaps by clicking on the document or by pressing `Alt`+`Tab`).
- In the word-processor document, place the cursor at the point just below your line of text. Choose **Paste** from the **Edit** menu of your word-processor (Alternatively, press `Ctrl`+`V`.)

Keep your word-processor document open as well as your Minitab session. You will need them both in the next couple of activities.

Computer activity 24 *Copying and pasting text from Minitab*

You can also copy text from any of the Minitab windows and insert it into a word-processor document. Insert the stemplot you obtained in Computer activity 19, into your word-processor document, by doing the following.

- In your word-processor document, add the following sentence below the cells copied from the worksheet: 'A stemplot of the distance between stops is as follows.'.
- Switch to your Minitab session and highlight the stemplot in your Session window.
- Select **Edit** and then **Copy** by clicking on it. (Alternatively, press `Ctrl`+`C`.)
- Now switch to your word-processor document.
- In the word-processor document, place the cursor at the point just below your line of text about the stemplot. Choose **Paste** from the **Edit** menu of your word-processor. (Alternatively, press `Ctrl`+`V`.)

Computer activity 25 *Copying and pasting graphical output from Minitab*

Graphs in Minitab can also be pasted into word-processor documents. Paste the scatterplot produced in Computer activity 19 into your word-processor document by doing the following.

- In your word-processor document, add the following sentence below the stemplot from the worksheet: 'A scatterplot of the miles per gallon (mpg) against distance between stops is as follows.'.
- Switch to your Minitab session and make the window containing the scatterplot the active window.
- Select **Edit** and then **Copy Graph** by clicking on it. (Alternatively, press `Ctrl`+`C`.)
- Now switch to your word-processor document.
- In the word-processor document, place the cursor at the point just below your line of text about the scatterplot. Choose **Paste** from the **Edit** menu of your word-processor. (Alternatively, press `Ctrl`+`V`.)

The word-processor document you were working with in Computer activities 23 to 25 is no longer needed, so you can close it now if you wish. However, if possible keep your Minitab session going until the end of Computer activity 27.

Computer activity 26 *Saving your session*

It is not always possible to do everything you want in a single Minitab session. In Minitab you can save your current session as a *project file*. This allows you to pick up where you left off at a later date. Do this now for your current session by doing the following.

- In Minitab, choose **Save Project** from the **File** menu.
- Select the folder where you would like to save the project file and enter the file name for the project into the dialogue box which opens. Project files have the extension **mpj**.
- Click on **Save**, or press `Enter`.

Minitab automatically adds **.mpj** to your file name.

If, at some later stage, you wish to return to Minitab and continue working on the project from where you left off, you should choose **Open Project** from the **File** menu, navigate to the correct location and enter the project name in the appropriate field. When you open the selected project file, you will find that the project is restored exactly as you left it, complete with worksheets, graphs and so on. If, after further work, you wish to save your session under a different name, choose **Save Project As...** from the **File** menu, and specify the new file name in the dialogue box that opens. If you choose **Save Project** instead, this will overwrite the old file with the new one.

Note that you can edit the Session window if you wish, for example to annotate it or remove unwanted text, before you save your session.

Computer activity 27 *Saving individual windows*

In Minitab it is also possible to save individual windows instead of the whole session. Save the window containing the scatterplot you obtained in Computer activity 19 by doing the following.

- Make the window containing the scatterplot active.
- Click on **File** and then choose **Save Graph As...**. (If you have already saved the window once, you will also be able to choose **Save Graph**. This saves the graph using its current file name and folder.)

The Session window and worksheets can be saved in a similar way. The only difference is that the dialogue boxes are obtained by clicking on **Save Session Window As...**, **Save Session Window**, **Save Current Worksheet As...** or **Save Current Worksheet**. (Note that the **Save Session Window** and **Save Current Worksheet** options may not be available if the Session window or worksheet have not been saved before.)

If you accidentally delete a data file, then you will need to obtain another copy from the CD.

Note that the M140 data files are read-only, so you will have to use **Save As...** rather than **Save** if you wish to save a worksheet that you have changed.

Summary of Chapter 1

In this chapter, you have used some of the features of Minitab to produce scatterplots of two datasets from Unit 1 and to do calculations on data using Minitab functions. In particular, you have learned how to round data to a specified number of decimal places.

You have learned how to obtain stemplots, including separate stemplots for subgroups of the data, and you have experimented with various ways to customise your stemplots, by trimming outliers and changing the number of parts each level is split into. You have also learned about histograms: how to obtain histograms using Minitab, and how to control where the 'bins' begin and end.

You have seen that Minitab stores data in files called worksheets, which have the file name extension **mtw**. You have learned how to print worksheets and output from Minitab, and how to copy worksheets and output into a word-processor document. You have also learned that sessions can be saved as projects, with the file name extension **mpj**. Projects store the worksheets, graphs and history of the session so that you can return to them later.

2 Measures of location

This chapter, which is associated with Unit 2, focuses on measures of location. In Subsection 2.1, you will be using an interactive computer resource to explore the resistance of such measures – that is, the impact, if any, that outliers have on the measure. Then, in Subsection 2.2, you will be exploring the properties of weighted means using a different interactive computer resource. (Study of this chapter does not involve the use of Minitab.)

Both subsections in this chapter are quite short, so you may find that you are able to complete their study in a single session.

2.1 Exploring measures of location

In Subsection 1.4 of Unit 2, you learned that the median is a resistant measure whereas the mean is a sensitive measure. Explore further what this means by completing the following activity.

Computer activity 28 *Resistance of the median and mean*

Open the interactive computer resource 'Sensitivity and resistance' on the M140 website.

This resource shows a plot of the data for prices of small televisions, introduced in Activity 1 of Unit 2 (Subsection 1.2). This plot is similar to a stemplot turned on its side. Each point on the plot represents the price of one small television (to the nearest £10). Where different televisions have the same price, the corresponding points are stacked on top of each other.

The median and mean for these data have been calculated, and these are shown on the plot using vertical dotted lines.

(a) For these data, which is bigger: the median or the mean? How large is the difference between them?

(b) Increase the price of the most expensive television (by moving the relevant point on the resource). What happens to the median and mean as you do this?

(c) Also increase the price of the second most expensive television. What happens to the median and mean now?

(d) How many of the prices of the most expensive televisions can be increased before the median changes?

2.2 Exploring weighted means

Subsection 2.1 of Unit 2 introduced the weighted mean – the mean of a combined batch – and the following three rules were stated.

Rule 1 The weighted mean depends on the relative sizes (i.e. the ratio) of the weights.

Rule 2 The weighted mean of two numbers always lies between the numbers and it is nearer the number that has the larger weight.

Rule 3 If the weights are equal, then the weighted mean of two numbers is the number halfway between them.

You are going to explore these rules by using an interactive computer resource in the next activity.

Computer activity 29 *Exploring weighted means*

In Example 8 of Unit 2 (Subsection 2.1), data about two batches of prices for small televisions were introduced. For batch A, with seven television prices, the mean was 119. For batch B, with 13 television prices, the mean was 185. Using this information, the mean of the combined batch was calculated.

A representation of this calculation is given in the interactive computer resource 'Weighted means'. Open this resource on the M140 website now.

In this representation of the data, the weights correspond to the sizes of the batches, and the positions of the weights correspond to the means of the batches.

(a) Move the fulcrum (pivot) to a position so that the bar is balanced horizontally. What does the position of the fulcrum now represent?

(b) In Example 8, the mean for the combined batch was calculated to be precisely 161.9. Move the fulcrum to this exact position. Does the bar balance?

(c) Double the weights on both sides of the balance bar. Where does the fulcrum now need to be for the bar to balance? How about when the weights on both sides are tripled?

(d) Now select new weights for the means by clicking on 'Random weights'. By finding the balance point, determine the weighted mean based on these new weights. Which batch mean is the weighted mean closer to? Is it less than 119 (the mean of batch A) or more than 185 (the mean of batch B)?

(e) Set both weights to be equal to 3. Where is the balance point now? Does it change when other pairs of equal weight are used?

Summary of Chapter 2

In this chapter, you have explored how the mean and median are affected by outliers – that is, how resistant or sensitive they are. You have seen that the median is resistant to changes in the data whereas the mean is sensitive. You have also explored the three rules associated with weighted means and seen that they hold true when applied to a particular dataset.

3 Summary measures and boxplots

This chapter, which is associated with Unit 3, focuses on summary measures and boxplots. In Subsection 3.1 you will use an interactive computer resource to compare the resistance of different measures of spread. Then, in Subsection 3.2, you will use Minitab to calculate measures of location and spread. Finally, in Subsection 3.3, you will learn how to produce boxplots using Minitab.

Subsection 3.1 is the shortest subsection and Subsection 3.3 is the longest subsection, so you should plan your time accordingly.

3.1 Exploring measures of spread

In Computer activity 28 you explored the resistance of the median and mean using the interactive computer resource 'Sensitivity and resistance'. In this short subsection, which consists of one activity, you will return to this interactive computer resource to explore the resistance of the range, interquartile range and standard deviation.

Computer activity 30 *Comparing the range, interquartile range and standard deviation*

Open the interactive computer resource 'Sensitivity and resistance' on the M140 website. Remember that the plot displays the prices of 20 small televisions. Start by selecting 'Show quartiles' and 'Show range'.

(a) Write down the range, interquartile range and standard deviation for the original data.

(b) What happens to the range, interquartile range and standard deviation when the most expensive small television is made even more expensive?

(c) What happens to the range, interquartile range and standard deviation when the cheapest television is made even cheaper?

(d) The median cost of a television is £150. What happens to the range, interquartile range and standard deviation when the price of one such average-priced small television is changed?

(e) Based on your findings in parts (b), (c) and (d), comment on the resistance/sensitivity of the range, interquartile range and standard deviation.

3.2 Calculating measures of location and spread using Minitab

In this subsection, you will be using Minitab to calculate all the measures of location, spread and so on that you have met in Units 1 to 3. Compared to the calculations in the units, this turns out to be far less tedious.

Run Minitab now, and open the worksheet **smallTVs.mtw**. This worksheet contains just one column, called `smallTVs`, containing the data on prices of small flat-screen televisions that you first met in Activity 1 of Unit 2 (Subsection 1.2). You probably worked quite hard in Units 2 and 3 on finding the median, quartiles and other measures based on these data. How could you have done all that in Minitab?

Computer activity 31 *Descriptive statistics in Minitab*

Minitab refers to quantities like the median, quartiles, mean, standard deviation and so on as *descriptive statistics*. Use Minitab to calculate descriptive statistics for the variable `smallTVs` by doing the following.

- Click on **Stat**, then **Basic Statistics**, then **Display Descriptive Statistics...**.
- The **Display Descriptive Statistics** dialogue box will appear. Enter the variable `smallTVs` in the **Variables** field. The completed dialogue box should be as in Figure 14.

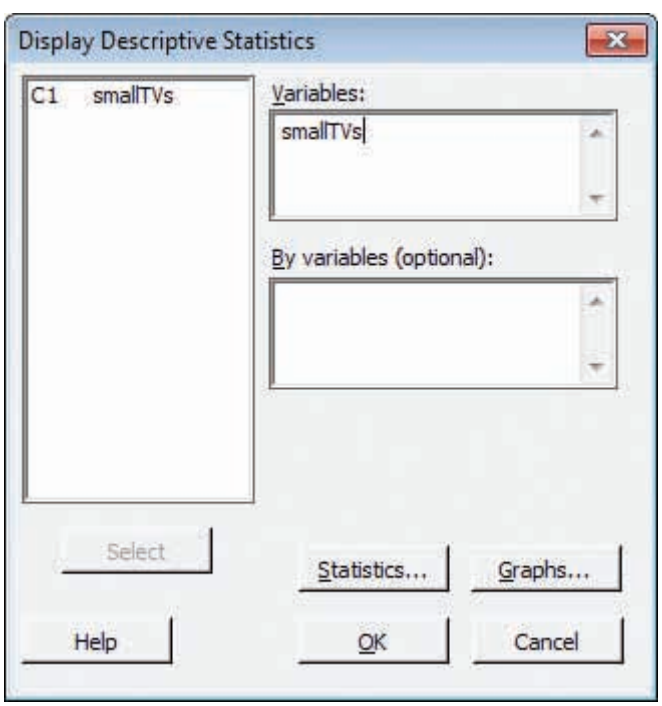


Figure 14 The **Display Descriptive Statistics** dialogue box

- Click on **OK**.

The following will appear in the Session window.

```
Variable   N  N*   Mean  SE Mean  StDev  Minimum     Q1  Median     Q3  Maximum
smallTVs  20   0  162.0     10.5   47.0     90.0  130.0   150.0  180.0    270.0
```

(a) Most (though not all) of the 'descriptive statistics' that you have already met in M140 are given in the output above, although a few are missing. Which of the quantities given in the output do you recognise?

(b) Other descriptive statistics that you have met in M140 can be calculated quite easily from the quantities Minitab has already given you: the interquartile range can be found from the quartiles, the range can be found from the extremes, and the variance can be found from the standard deviation. (The standard deviation is the square root of the variance, so the variance is the standard deviation squared.) Use your calculator to find the range, interquartile range and variance of the `smallTVs` data, using the appropriate values that were given by Minitab.

Computer activity 32 *Customising Minitab's descriptive statistics*

In Computer activity 31 you obtained the range, interquartile range and variance by calculating them from other descriptive statistics given by Minitab. But why should you get your calculator out if you can make Minitab do the work for you?

- Make sure that **smallTVs.mtw** is the active window in Minitab.
- Obtain the **Display Descriptive Statistics** dialogue box again (**Stat** > **Basic Statistics** > **Display Descriptive Statistics**).
- You will probably find that `smallTVs` is already in the **Variables** field; if not, enter it there again. This time, instead of clicking on **OK**, click on **Statistics...**.
- A new dialogue box appears, called **Display Descriptive Statistics: Statistics**. This has a lot of names of descriptive statistics, each with a tick box beside it. If the box is ticked, the corresponding statistic will be displayed in the final output. You can change whether an item is ticked by clicking on its name or its tick box.

  Click on the boxes for the range, the interquartile range and the variance to select them. Also click on the boxes for 'SE of mean' and 'N missing' to deselect them. (If you like, you can deselect some of the other boxes that were selected, since you already have the values of those statistics.) The resulting dialogue box is shown in Figure 15.

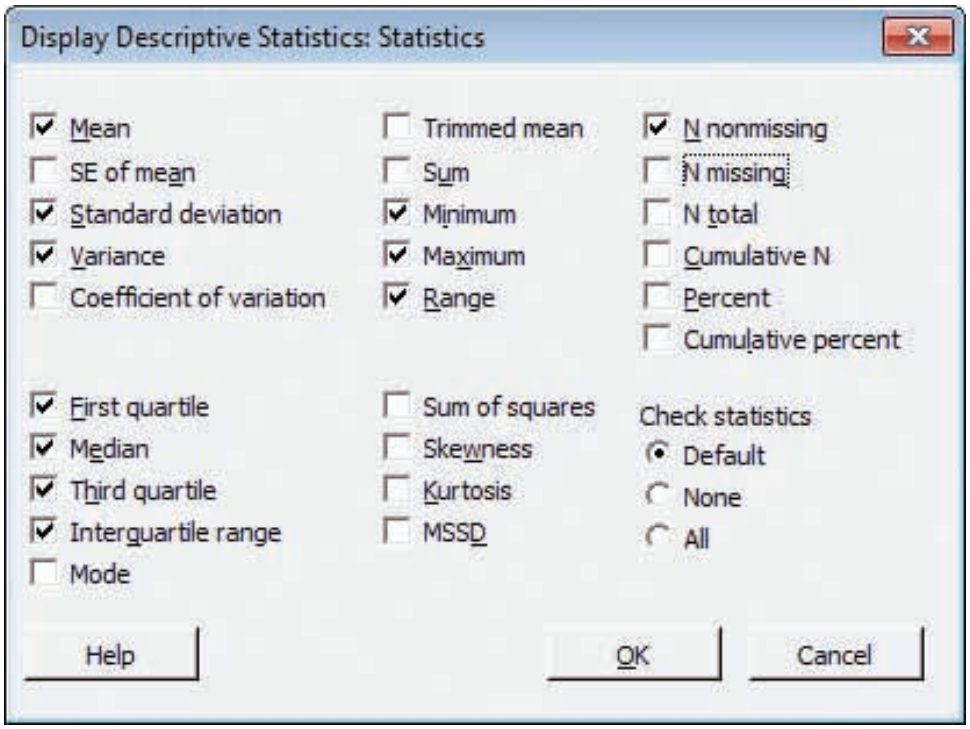


Figure 15 The **Display Descriptive Statistics: Statistics** dialogue box

- Now click on **OK**, and then click on **OK** in the main **Display Descriptive Statistics** dialogue box.

Are the range, interquartile range and variance the same as you found in Computer activity 31?

Computer activity 33 *Comparing batches with Minitab's descriptive statistics*

Minitab will calculate descriptive statistics for more than one batch of data at a time, and display them in a way that makes it relatively easy to compare the batches.

Open the worksheet **pay.mtw** and make sure it is the active window. This contains the data from Activity 13 of Unit 3 (Subsection 2.2) on the hourly earnings of 40 women, along with the data summarised in Exercise 4 of Unit 3 (Exercises on Section 2) for 35 men. (In the worksheet, the data are given to the nearest penny, whereas in the unit, they were truncated to whole pounds, so the calculations from Minitab will give slightly different results to those in the unit.)

- Obtain the **Display Descriptive Statistics** dialogue box again (**Stat** > **Basic Statistics** > **Display Descriptive Statistics**).
- This time, enter both `women` and `men` in the **Variables** field. (You might have to delete `smallTVs` from that field.)

- Before you click on **OK**, click on **Statistics...** to give the **Display Descriptive Statistics: Statistics** dialogue box. Make sure that **Mean**, **Standard deviation**, **Minimum**, **Maximum**, **N nonmissing**, **First quartile**, **Median**, **Third quartile** and **Interquartile range** are selected.
- Click on **OK** to get back to the main dialogue box, and then on **OK** in that dialogue box.

The display shown below will appear in the Session window.

Variable	N	Mean	StDev	Minimum	Q1	Median	Q3	Maximum	IQR
women	40	13.87	9.16	6.05	8.03	10.29	16.89	43.14	8.86
men	35	14.38	7.15	6.52	8.49	12.79	17.75	38.17	9.26

This allows you to see and compare the descriptive statistics at a glance. For instance, the mean hourly earnings is higher for men than for women. However, the position on spread is not so clear – the standard deviation is higher for the women than the men, but the interquartile range is higher for the men than the women. (This is because of the differing resistance of the standard deviation and the interquartile range to changes in the extremes – both these batches contain some particularly high values.)

3.3 Boxplots in Minitab

In this subsection, you will learn how to use Minitab to create boxplots – a reasonably straightforward task.

Computer activity 34 *Drawing a boxplot in Minitab*

There are actually several different ways of producing a boxplot in Minitab. The one you will use in this activity uses the **Boxplot...** command on the **Graph** menu. With **smallTVs.mtw** as the active window in Minitab, do the following.

- Select **Graph** and then **Boxplot...**.
- The **Boxplots** dialogue box should appear. This has four small pictures: **One Y, Simple**; **One Y, With Groups**; **Multiple Y's, Simple**; and **Multiple Y's, With Groups**. We will be using only the **Simple** versions in M140. Make sure that **One Y, Simple** is selected, and click on **OK**.
- The **Boxplot: One Y, Simple** dialogue box appears. Copy the variable `smallTVs` into the **Graph variables** field. The completed dialogue box should be the same as in Figure 16.

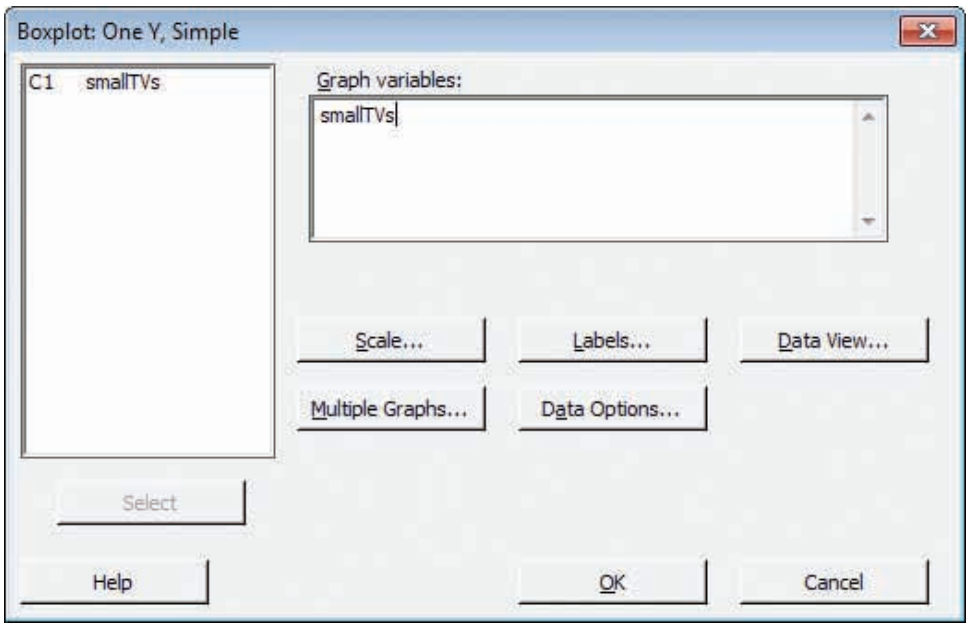


Figure 16 The **Boxplot: One Y, Simple** dialogue box

- Click on **OK**.

A new window will open, containing a boxplot. It is shown in Figure 17.

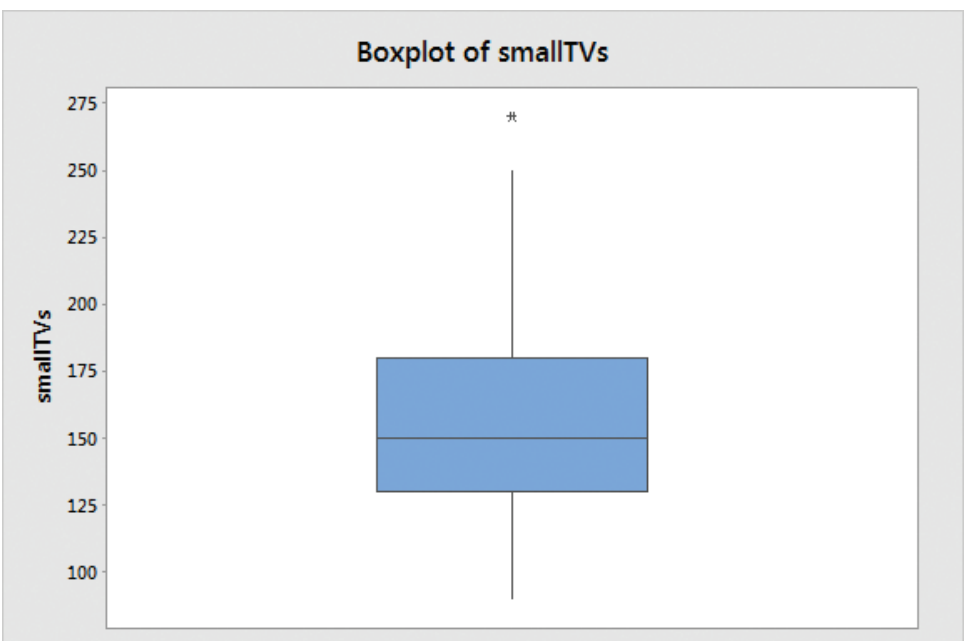


Figure 17 A boxplot drawn by Minitab

Minitab has drawn the boxplot vertically – its default. However, you can change it to produce a horizontal boxplot, as we have been drawing in the units.

Obtain the **Boxplot: One Y, Simple** dialogue box again (**Graph** > **Boxplot** > **One Y, Simple**). You will have noticed that it has various buttons in its middle section, and these can be used to customise the plots in various ways.

To make the plot horizontal, do the following.

- Click on **Scale...**. This produces the **Boxplot: Scale** dialogue box.
- Click on the **Transpose value and category scales** tick box to select it. This will have the effect of making the boxplot horizontal.
- Click on **OK**, and then click on **OK** in the main **Boxplot: One Y, Simple** dialogue box.

This should produce the boxplot in Figure 18, which looks rather more like those you have seen in Units 2 and 3.

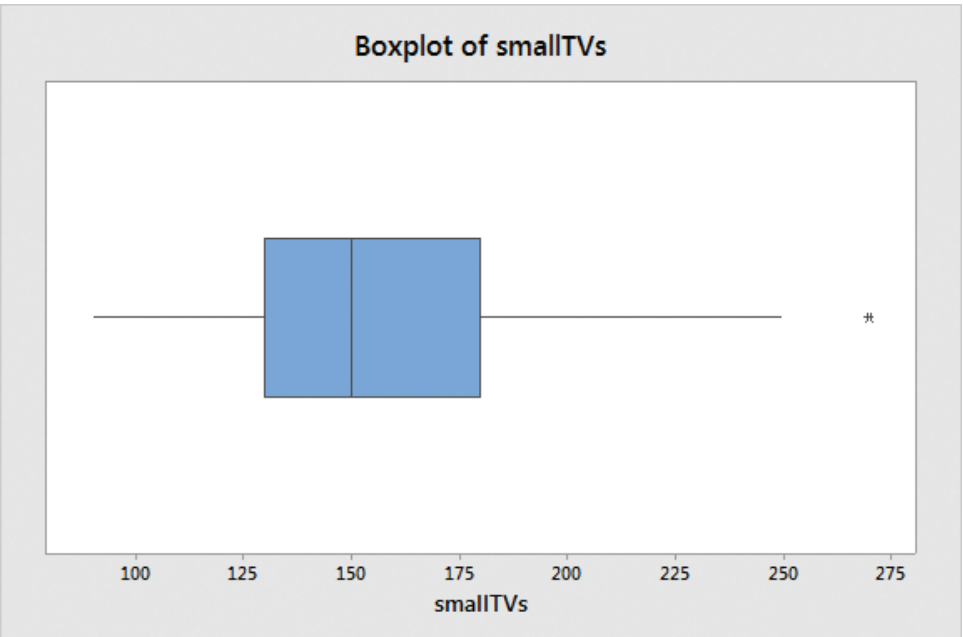


Figure 18 A horizontal boxplot drawn using Minitab

If possible, keep the window containing this boxplot open in Minitab. You will need it in the next activity.

Computer activity 35 *Customising a boxplot*

The boxplots in Computer activity 34 were labelled using the variable name `smallTVs`. This is good enough when it is just you who is looking at the boxplot, as you know what the variable `smallTVs` represents. However, better labelling is required when others are going to be looking at the boxplot as well – for example, when you include the boxplot in a report or TMA.

Alter the labelling on the horizontal boxplot you produced in Computer activity 34 by doing the following.

- Make the window containing the horizontal boxplot the active window. (If you do not have this window open in Minitab, then follow the instructions in Computer activity 34 to recreate it.)
- Double-click on the title of the boxplot. This brings up the **Edit Title** dialogue box.

  Alternatively, click on **Editor** and then **Select item** to bring up a submenu listing the items on the boxplot that can be selected. You are going to alter the title, so click on **Title: Boxplot of smallTVs**. Notice that marks are now placed around the boxplot title to indicate it has been selected. Then click on **Editor** again and click on **Edit Title: Boxplot of smallTVs...**. This brings up the **Edit Title** dialogue box.
- In the **Edit Title** dialogue box, the text in the **Text** field will be used for the title. Change this text to the following more informative title: `Boxplot showing the price variation in small TVs`.
- Click on **OK**.

Notice that the title of the boxplot has now been changed to match the amended text above.

The label for the horizontal axis also requires changing to something more informative. This can be done in a similar way to changing the title, as follows.

- After checking that the boxplot window is the active window, bring up the **Edit Axis Label** dialogue box by double-clicking on the axis label `smallTVs`.

  Alternatively, click on **Editor** then **Select Item** and then **Y axis label**. Notice that, this time, markers are placed round the label `smallTVs` on the horizontal axis. Click on **Editor** and then **Edit Y axis label**.
- In the **Edit Axis label** dialogue box, change the **Text** field so that it contains `Price (to the nearest £10)`.
- Click on **OK**.

The boxplot should now look like that in Figure 19.

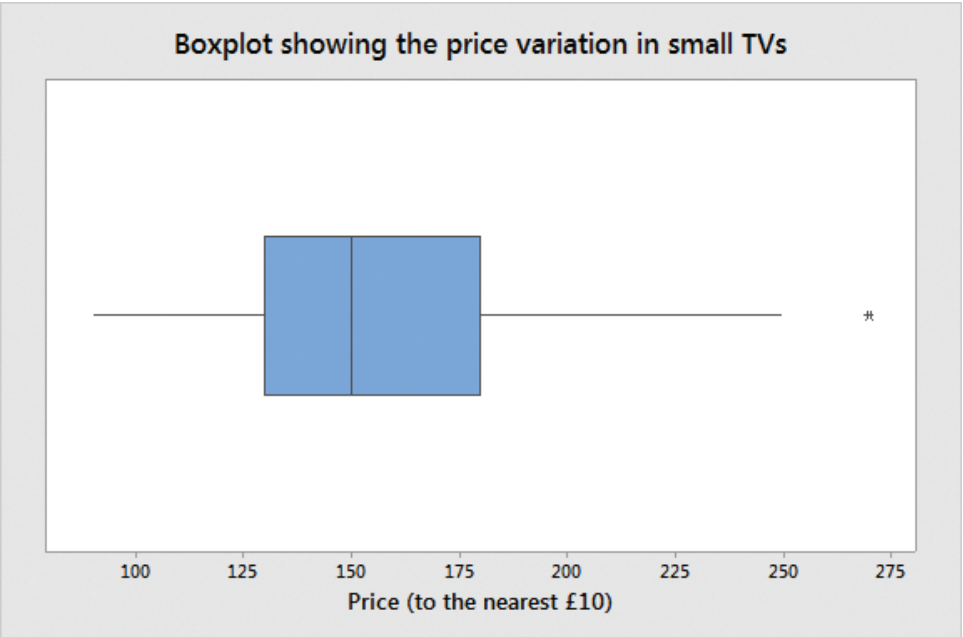


Figure 19 An appropriately labelled boxplot

In Computer activity 35, you altered the title and axis label of a boxplot to something more informative than Minitab automatically provided. The same strategy can be used to customise other sorts of plot, such as the scatterplot you produced in Computer activity 4. Just double-click on the part of the plot you want to alter to bring up the corresponding **Edit** dialogue box. (Or use **Editor** > **Select item** to select the part of the plot you want to alter. Then select **Editor** > **Edit**.)

On a plot, the features that can be altered include symbols, line styles and line thicknesses. You can also alter the size of the plot. The goal is to produce a plot that is clear and not misleading, rather than one that is aesthetically pleasing, and sometimes this may involve changing the default plot produced by Minitab.

Computer activity 36 *Multiple boxplots in Minitab*

You saw several examples in Unit 3 where boxplots from more than one batch of data were plotted on the same diagram. This made it relatively easy to compare the batches. This activity shows you how it is done in Minitab.

- With **pay.mtw** as the active window, select the **Boxplot...** command on the **Graph** menu.
- This time, in the resulting **Boxplots** dialogue box, choose **Multiple Y's, Simple** and click on **OK**. (The 'multiple' is because we are going to plot the boxplots of women's and men's hourly earnings on the same diagram.)

- Unsurprisingly, this opens the **Boxplot: Multiple Y's, Simple** dialogue box. In fact, apart from its title, this looks just like the **Boxplot: One Y, Simple** dialogue box. Enter both `women` and `men` into the **Graph variables** field.
- To produce horizontal plots, click on **Scale...** to obtain the **Boxplot: Scale** dialogue box, and click on the **Transpose value and category scales** tick box to select it.
- Click on **OK**, and then click on **OK** in the main **Boxplot: Multiple Y's, Simple** dialogue box.

The resulting diagram, containing both boxplots, is shown in Figure 20.

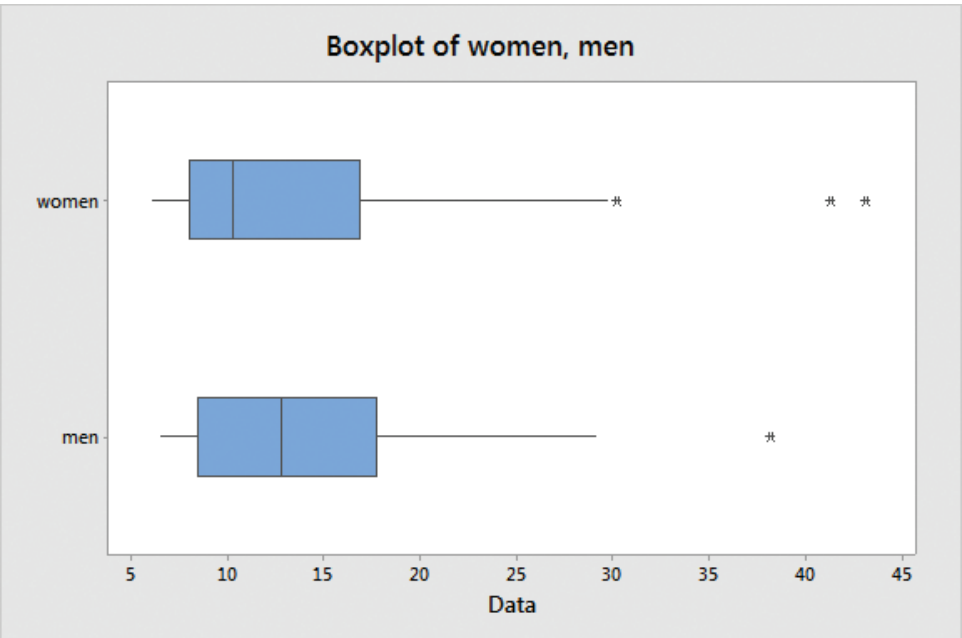


Figure 20 Boxplots of men's and women's pay

From these boxplots it is possible to see that generally the men's earnings are slightly higher at the key values, though women's earnings are higher at the upper extreme. Both batches generally look right-skew, but the women's batch is rather more skewed.

Computer activity 37 *Comparing times from a 10 km race*

The worksheet **times10k.mtw** contains the times that competitors took to complete a 10 km race in May 2012. (The times are given in minutes in decimal format. So, for example, 45.5 corresponds to a time of 45 minutes and 30 seconds.)

There are four variables given in the worksheet:

- `senior_men` – times for senior men (ages 21 to 39)
- `veteran_men` – times for men in a veteran class (ages 40 to 44)
- `senior_women` – times for senior women (ages 21 to 39)
- `veteran_women` – times for women in a veteran class (ages 40 to 44).

(a) Use Minitab to create a single diagram containing horizontal boxplots of the data in **times10k.mtw**.

(b) Change the title and axis labelling of the diagram to be more informative.

(c) Use the boxplots to compare the times achieved by men and women.

(d) Use the boxplots to compare the times achieved by the senior and veteran competitors.

Summary of Chapter 3

In this chapter, you have learned how resistant or sensitive the measures of spread are – that is, how they are affected by changes in the data, especially changes to the extreme values. In particular, you have learned that the interquartile range is resistant and the standard deviation and range are sensitive.

You have learned how to use Minitab to calculate quantities such as the mean, median, quartiles and standard deviation, and you have seen how to customise the quantities Minitab produces. You have also learned how to produce boxplots (single and multiple) using Minitab and how to customise boxplots by changing their orientation and labelling.

4 Sampling

This chapter, which is associated with Unit 4, focuses on sampling. In Subsection 4.1, which is quite short, you will use an interactive computer resource to explore a sampling distribution. Then, in Subsection 4.2, you will use Minitab to generate some simple random samples.

4.1 Exploring a sampling distribution

Subsection 3.3 of Unit 4 introduced the notion of a sampling distribution. In the following activities you will explore properties of some sampling distributions using an interactive computer resource.

Computer activity 38 *Properties of a sample of size 3*

Open the interactive computer resource 'Sampling distribution of the median' on the M140 website. This resource uses the same target population as Subsection 3.2 of Unit 4 – that is, a population of 1000 individuals whose responses to the following question are given in the following table.

> *Considering what has happened to your earnings, the way prices have changed and changes in other circumstances, do you feel that you are now better or worse off than you were twelve months ago?*

Population values of the response

Response	Rating	Number
Much worse off	1	300
Somewhat worse off	2	100
About the same	3	200
Somewhat better off	4	300
Much better off	5	100

(a) Generate a sample of size 3 from this population by following the instructions in the resource.

A bar will appear on the graph, labelled with a '1' to indicate that one sample has been taken so far. The x-axis indicates the median response for the sample, and the median is also displayed beneath the plot – along with the values in the sample.

What values do you get in your sample? What is the median response for your sample? Does it match the population median response?

(b) Take nine further samples of size 3, by clicking on 'Generate sample' a further nine times (taking your 'Sample count' to 10).

Each time you generate a sample, a new bar will be added or the frequency of an existing bar will increase. The median and values of the last sample generated will be displayed beneath the graph, along with a count of the number of samples taken.

How often does the median response match the population median?

(c) Take 100 more samples of size 3, by clicking on 'Generate 100 samples' (taking your 'Sample count' to 110). Use the results to complete the following table.

Sample median response	Number of samples	Proportion of samples
1		
2		
3		
4		
5		
Total		

(d) Compare your results in part (c) with the proportions that are based on all samples. (The latter are given in Table 10 of Unit 4 (Subsection 3.3).)

Computer activity 39 *Changing the sample size*

In Computer activity 38, you explored the sampling distribution for samples of size 3 from a population. In this activity, you will explore the sampling distributions associated with samples of different sizes. You will consider only sample sizes that are odd numbers, so that the sample median is always equal to a sampled data value.

- Open the interactive computer resource 'Sampling distribution of the median' if it is not already open.
- Click on 'Reset'.

In Computer activity 38, you saw that only about a third of samples of size 3 had the same median as the population.

(a) With the 'Sample size' set to `3`, generate 100 samples. What proportion has a median of 3?

(b) Click on 'Reset'. Change the 'Sample size' to `5` and generate 100 samples. What proportion has a median of 3?

(c) Click on 'Reset'. Change the 'Sample size' to `11` and generate 100 samples. What proportion has a median of 3?

(d) By generating 100 samples each of sizes 21, 41 and 81, complete the following table.

Sample size	Number of samples with sample median 3
3	*your answer from (a)*
5	*your answer from (b)*
11	*your answer from (c)*
21	
41	
81	

Hence suggest how big a sample should be taken so that more than 90% of such samples will have a sample median of 3.

In Computer activity 38 you saw that for one population, when the sample size is just 3, many – if not most – of the samples did not contain the population median. However, as you have seen in Computer activity 39, as the sample size increases, the sample median becomes much more predictable, and more likely to be equal to the population median.

4.2 Generating simple random samples

In this subsection, you will learn how to use Minitab to generate simple random samples. Computer activity 41 uses a worksheet created in Computer activity 40, so you should try to do these two activities in the same study session.

Computer activity 40 *Random sample from a population of 100*

In this activity you will get Minitab to choose labels for a random sample of size 15 from a population of 100 individuals.

- Open a new Minitab worksheet, by clicking on **File** and then **New**. In the **New** dialogue box select 'Minitab Worksheet' before clicking on **OK**. (Or, if you have just opened Minitab, you can use the new worksheet that will be open already.)
- Click on **Calc** and then open the **Random Data** submenu.
- You are going to generate integer values, so select **Integer...** from the list that is now displayed.
- The **Integer Distribution** dialogue box will appear. In this dialogue box you are asked to specify the number of rows of data you wish to generate, the column in which you want to store the data, a minimum value and a maximum value. Specify these now by doing the following.
  - In the **Number of rows of data to generate** field, type in `30`. (This will generate 30 random numbers to allow us to discard any repeated values.)
  - In the field called **Store in column(s)**, type in `C1`.
  - Enter a minimum value of `1` and a maximum value of `100`. (This will generate numbers between 1 and 100 inclusive.)

  The completed dialogue box should be the same as Figure 21.

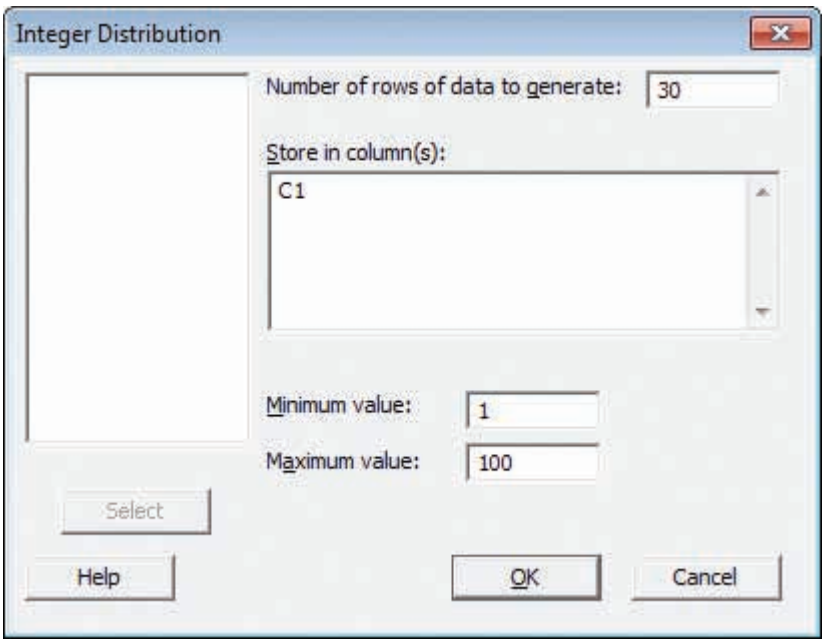


Figure 21 The **Integer Distribution** dialogue box

- Click on **OK**.

The 30 random numbers generated will appear in column `C1` of the Minitab worksheet.

(a) Which random numbers did Minitab select?

(b) Were any random numbers selected more than once? If so, which?

(c) In Subsection 1.2 of Unit 4, the simple random sample corresponded to the random numbers in the order in which they occurred along a row of a random number table, ignoring any repeated random numbers. Applying the same rule here – 'take the numbers in the order in which they are generated, ignoring repeats' – which random numbers constitute a random sample of 15 from a population of 100?

If possible, leave this worksheet open in Minitab. You will need it for the next activity.

With a small sample, it is fairly easy to spot repeated values; this is not so easy with a larger sample. We can, however, use Minitab to tabulate the random numbers to help with this task.

Computer activity 41 *Tabulating random numbers*

If you closed the worksheet that you used for Computer activity 40, then you will need to generate another set of 30 random numbers before continuing. Use Minitab to tabulate the random numbers by doing the following.

- Click on **Stat**, then **Tables**, and then **Tally Individual Variables...**.
- In the **Tally Individual Variables** dialogue box that appears, enter `C1` in the **Variables** field. Also make sure that the option **Counts** is selected. The completed dialogue box should be the same as in Figure 22.

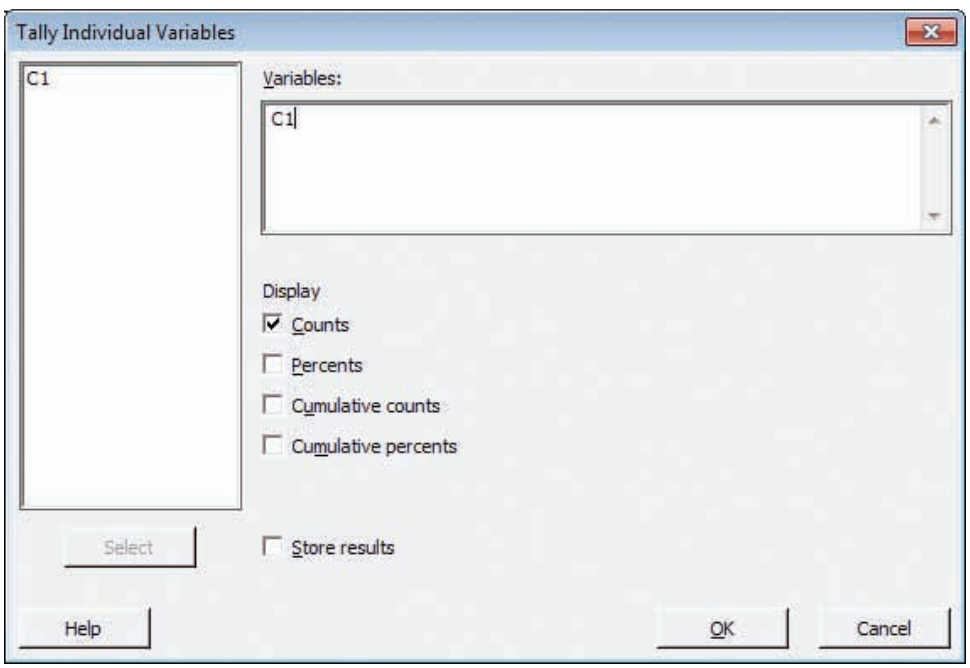


Figure 22 The **Tally Individual Variables** dialogue box

- Click on **OK**.

A list of all the different values in column `C1` is then displayed in the Session window, along with the number of times each value occurs. For example, for the sample given in the solution to Computer activity 40(a), the following list is given.

C1	Count
1	1
6	2
9	1
11	1
12	1
27	1
28	1
36	2
40	1
41	1
43	3
46	1
49	2
50	2
53	1
54	1
59	1
62	1
67	1
73	1
75	1
83	1
98	2
N=	30

So it is easy to see that in this sample the value 6 occurs twice whereas the value 9 occurs just once. Values that are not listed in the first column (such as 2, 3, 4 and 5) do not occur at all.

In column `C1`, delete any repeated values identified in this list by clicking in the appropriate cell of the worksheet and deleting the relevant value. The deleted number will be replaced by $*$.

The first 15 values left in `C1` are labels for the 15 individuals to be selected from the population of 100.

Computer activity 42 *Random sample from a population of 1000*

Use Minitab to choose a random sample of size 10 from a population of size 1000.

In Computer activities 40 and 41 you generated a random sample by generating a set of random numbers and removing duplicates. When the population is relatively small compared to the sample required, this process can get tedious – many duplicates may have to be identified and removed. In the following two activities you will learn an alternative way

of using Minitab to obtain a simple random sample. In Computer activity 43 you will create a Minitab worksheet in which labels for every single member of the population are given in the first column. Then, in Computer activity 44, you will get Minitab to select a simple random sample from a population given in a worksheet. Computer activity 44 uses the worksheet that you will produce in Computer activity 43, so you should try to do these two activities in the same study session if possible.

Computer activity 43 *Entering labels for a population*

In Computer activity 40, the labels $1, 2, 3, \ldots, 100$ were used to represent a population of 100. Create a Minitab worksheet in which all these labels are given in the first column by doing the following.

- Open a new worksheet in Minitab (**File** > **New**).
- Click on **Calc**, then **Make Patterned Data**, and then click on **Simple Set of Numbers...**. The **Simple Set of Numbers** dialogue box will open.
- Suppose the column containing the population labels is to be called 'population'. Type `population` in the field labelled **Store patterned data in**.
- Type `1` in the **From first value** field and `100` in the **To last value** field. Make sure that `1` is entered in the other three fields. The dialogue box should now look like Figure 23.

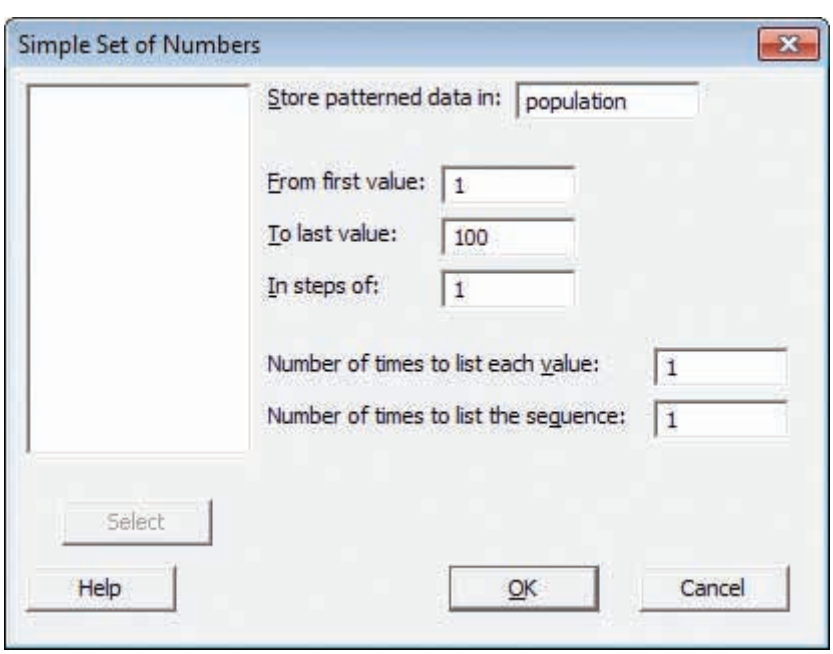


Figure 23 The **Simple Set of Numbers** dialogue box

- Click on **OK**.

The worksheet will now contain the numbers `1`, `2`, ..., `100` in a column labelled `population`. Continue straight on to Computer activity 44 as you will need this worksheet in that activity.

Computer activity 44 *Sampling from a population*

In the previous activity, you entered all the labels for a population in a Minitab worksheet. In this activity you will get Minitab to take a simple random sample of size 15 from this population of 100.

- Make sure that the worksheet you created in Computer activity 43 is the active worksheet in Minitab.
- Click on **Calc**, then on **Random Data**, and then on **Sample From Columns...**.
- In the **Sample From Columns** dialogue box, enter `15` in the **Number of rows to sample** field. In the **From columns** field enter `population`, and in the **Store samples in** field enter `sample`. Make sure that **Sample with replacement** is not selected. The dialogue box should now look like Figure 24.

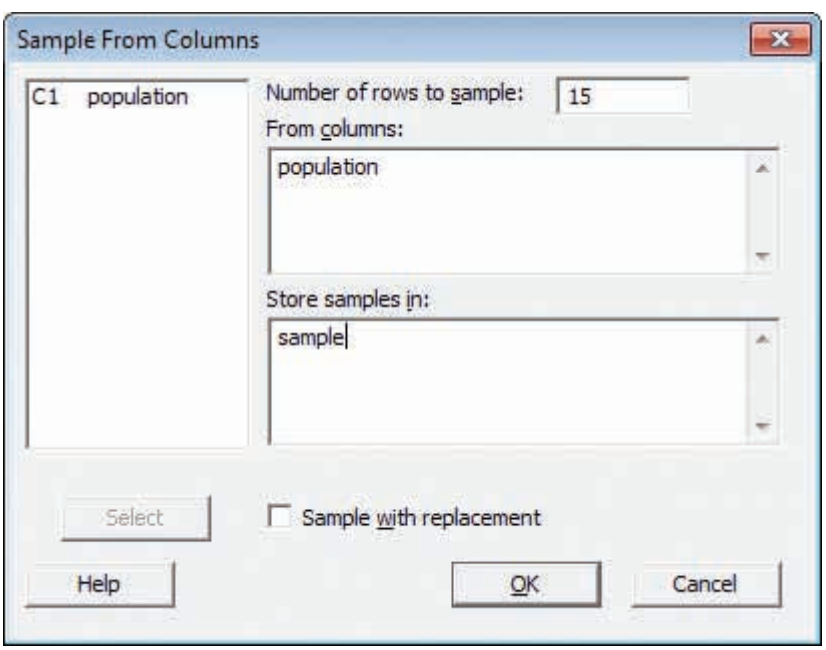


Figure 24 The **Sample From Columns** dialogue box

- Click on **OK**.

A simple random sample of size 15 from the labels given in the column `population` is then given in the column `sample`. Note that the sample automatically does not contain any duplicates as the **Sample with replacement** box was not selected.

Computer activity 45 *Random sample from a population of 86*

(a) Use Minitab to choose a random sample of size 15 from the Sampling Department of a large organisation. The staff list for the department is given in Table 1 of Unit 4 (Section 2).

(b) Note the gender and occupation of each individual selected. By comparing the information from your sample with the percentages of department staff by gender and occupation given in Table 4 of Unit 4 (Subsection 2.1), comment on the representativeness of the sample with respect to gender and occupation.

Summary of Chapter 4

In this chapter, you have explored the distribution of a sample median. You have seen that the sample median is more likely to be equal to the population median as the sample size increases.

You have learned how to use Minitab to obtain simple random samples in two ways. One way is to use Minitab to generate random numbers from a range corresponding to the size of the population, discarding duplicated numbers until the sample size is reached. This has the advantage of not having to enter all the labels for the population into Minitab first. The other way is to enter all the labels for the population in a Minitab worksheet and get Minitab to select a simple random sample of those labels. This has the advantage of not having to check for duplicate random numbers.

5 Relationships

This chapter, which is associated with Unit 5, focuses on relationships in data. In Subsection 5.1 you will use a couple of interactive computer resources to explore the fitting of least squares regression lines. In Subsection 5.2 you will learn how to calculate the equation of the least squares regression line using Minitab. Also in this subsection, you will learn how to use Minitab to generate scatterplots that display the least squares regression line. Then, in Subsection 5.3, you will learn how to produce the associated residual plot.

5.1 Fitting lines

In Subsection 4.1 of Unit 5 it was stated that the least squares regression line goes through the point $(\overline{x}, \overline{y})$ and minimises the sum of the squares of the residuals. By completing the following couple of activities, explore some of the reasoning behind these criteria for fitting a line.

Computer activity 46 *Fitting lines*

Open the interactive computer resource 'Fitting lines'. This resource displays a scatterplot of 10 data points, along with a line. (This line does not necessarily represent a good fit to the 10 data points.) The corresponding residuals are shown, along with the sum of these residuals.

(a) Move the line up and down on the plot. Does the sum of the residuals increase when you move the line up or down?

(b) Move the line to a position where the sum of the residuals is zero. Does the point $(\overline{x}, \overline{y})$ lie on the line?

(c) Repeat part (b) a couple more times, each time starting with a new line. Do you get the same answer?

As you saw in Computer activity 46, making sure the fit line goes through $(\overline{x}, \overline{y})$ is a sensible requirement as it ensures the sum of the residuals is zero. However, it does not help determine the best slope to use. For this we need another criterion.

Computer activity 47 *Choosing a line to minimise the sum of the squared residuals*

Open the interactive computer resource 'Fitting lines using least squares'. Similarly to 'Fitting lines', this resource displays a scatterplot of 10 data points along with a line. The line is fixed to always go through the point $(\overline{x}, \overline{y})$, although it does not necessarily represent a good fit to the data. The plot also shows the residuals associated with the data points and line, along with the sum of the squared residuals.

(a) Make the line steeper. What do you notice about the sum of the squared residuals?

(b) Now make the line less steep. What do you notice about the sum of the squared residuals this time?

(c) Rotate the line so that the sum of the squared residuals is as small as possible. What is the equation of this line? Does this line appear to fit the data well?

(d) Display the least squares regression line by clicking on the relevant tick box of the resource. Does the position of this line match the position of the line you settled on in part (c)?

5.2 Calculating a least squares regression line

In this subsection you will learn how to calculate a least squares regression line using Minitab.

Computer activity 48 *Fitting a least squares regression line*

In Activity 17 of Unit 5 (Subsection 4.2), you calculated by hand the least squares line for the blood pressure data. In this activity, you will use Minitab to accomplish the same task. These data are given in the worksheet **captopril.mtw**. Open this worksheet in Minitab now.

- Click on **Stat**, click on **Regression**, then on **Regression** and choose **Fit Regression Model...**. The **Regression** dialogue box will open.

  For these data, the y values correspond to variable `after` and the x values correspond to variable `before`.

- Enter `after` in **Responses** and `before` in **Continuous predictors**. The completed dialogue box should be the same as in Figure 25.

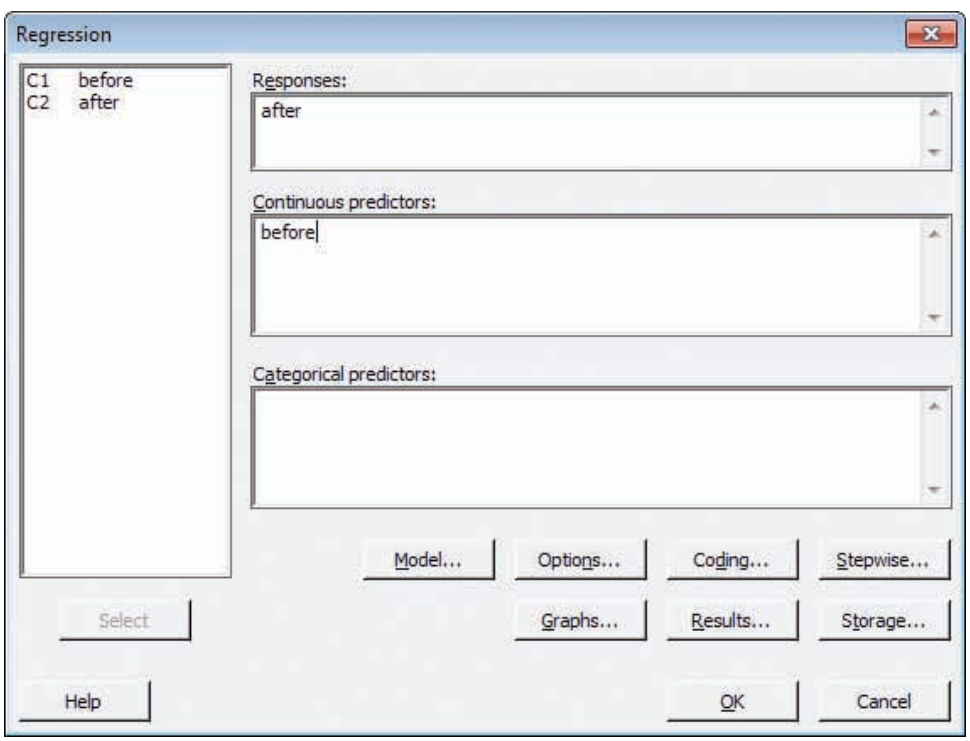


Figure 25 The **Regression** dialogue box

- Click on **OK**.

Look in the Session window. The equation of the fitted line, the 'regression equation', is given just after the heading. In the Data window a small marker has been added to the `after` column to show this is the response variable.

(a) Write down the equation of the least squares line calculated by Minitab.

(b) Compare the equation of the line given by Minitab with that calculated by hand in Activity 17 of Unit 5. Do they match?

Computer activity 49 *Adding a least squares regression line to a scatterplot*

In Computer activity 48 you used Minitab to obtain the least squares regression line. In this activity, you will add the least squares regression line to a scatterplot. When doing this, Minitab automatically calculates the equation of the least squares regression line. So it is not necessary to have done that using the procedure described in Computer activity 48 first.

You will now add the least squares regression line to the blood pressure data given in the worksheet **captopril.mtw**. Make this worksheet the active worksheet in Minitab now if it is not already.

- Produce a scatterplot of the data, with the blood pressure measurements after treatment on the y-axis. (**Graph** > **Scatterplot** > **Simple**, entering `after` in the **Y variables** field and `before` in the **X variables** field.)
- Make the Graph window the active window (**Window** > **Scatterplot of after vs before**).
- Click on **Editor**, then **Add**, then **Regression Fit...**. The **Add Regression Fit** dialogue box will open.
- Select 'Linear' as the **Model Order** and make sure that **Fit intercept** is selected. Click on **OK**.

Compare the scatterplot with that given in the solution to Activity 17 of Unit 5 (Subsection 4.2).

Computer activity 50 *Modelling international student achievement*

Activity 8 of Unit 5 (Subsection 2.3) introduced data on student achievement in different countries. The file **pisa.mtw** contains a copy of these data.

(a) Obtain a scatterplot with achievement on the reading scale along the x-axis and achievement on the mathematics scale along the y-axis.

(b) Describe the relationship between student achievement on the reading and mathematics scales.

(c) Obtain the equation of the least squares regression line.

(d) Display the least squares regression line on the scatterplot. Does the line appear to provide a reasonably good fit?

5.3 Residual plots

In this subsection you will learn how to produce residual plots for a least squares regression line.

Computer activity 51 *Producing a residual plot*

In Subsection 5.1 of Unit 5 you saw that residual plots should be used to check the fit of a least squares regression line to data. In this activity, you will learn to produce such residual plots using Minitab. As you will see, this is achieved without having to explicitly calculate the residuals first.

You will reproduce the residual plot given in Example 19 of Unit 5 (Subsection 5.1) – that is, the residual plot from the least squares fit line for the data on male unemployment and car ownership. These data are given in the worksheet **unemployment and cars.mtw**. Open this worksheet in Minitab now.

You can use Minitab to produce various kinds of residual plot at the same time as you use it to calculate the least squares regression line. So the first step is to set up the dialogue box ready to fit the least squares regression line.

- Click on **Stat**, click on **Regression**, then on **Regression**, and then on **Fit Regression model...**, to open up the **Regression** dialogue box.
- Enter `car` in **Responses** and `unemploy` in **Continuous predictors**.
- Click on **Graphs...** to open up the **Regression: Graphs** dialogue box.

There are different types of plot that can be used to check the fit of least squares regression lines. Here we will focus on just one, a scatterplot with the residuals on the y-axis and the explanatory variable, `unemploy`, on the x-axis.

- Make sure that the `Regular` option is selected under **Residuals for Plots**. In the **Residuals versus the variables** field, enter `unemploy`.
- Click on **OK** to return to the **Regression** dialogue box, and click on **OK**.

Compare the residual plot with that given in Figure 49 of Unit 5 (Subsection 5.1). Do they match?

Computer activity 52 *Checking the model for international student achievement*

In Computer activity 50, you fitted a least squares regression line to data on student achievement.

(a) Produce the corresponding residual plot.

(b) What does this residual plot indicate about the fit of the least squares regression line? Is it reasonable? Justify your opinion.

Summary of Chapter 5

In this chapter, you have experimented with fitting lines to data based on the sum of the residuals and the sum of the squared residuals, and you have seen that minimising the sum of the squared residuals is a reasonable criterion for fitting a line automatically. You have learned how to use Minitab to calculate the equation of a least squares regression line, i.e. the line which minimises the sum of the squared residuals, and how to display this regression line on a scatterplot. You have also learned how to produce residual plots using Minitab.

6 Probabilities and the sign test

This chapter accompanies Unit 6, where you learned (among other things) how to calculate probabilities for the number of items that will exceed the population median in a random sample, and how to perform a sign test.

The focus of this chapter is learning to use Minitab to do these tasks. In Subsection 6.1 you will learn how to calculate the probabilities, and in Subsection 6.2 you will learn how to perform the sign test using Minitab.

Note that Subsection 6.1 is longer than Subsection 6.2 and hence is likely to take longer to complete.

6.1 Calculating probabilities for the sign test

In Section 3 of Unit 6, you considered taking a random sample of size n and learned that the probability that exactly x of these observations are greater than the population median is

$$ {}^{n}C_{x} \times \left(\tfrac{1}{2}\right)^{n}. $$

Depending on your calculator, calculating these probabilities can be laborious unless n is small.

In the following activity, you will calculate such probabilities using Minitab. As mentioned briefly in Subsection 3.2 of Unit 6, the formula for the probabilities is a special case of a probability distribution called the binomial distribution. Hence 'binomial distribution' underlies the names of some of the Minitab commands you will use. You will need the results from Computer activity 53 for Computer activity 54, so you should do both these activities in the same session.

Computer activity 53 *Probability distribution for a random sample of size 12*

Subsection 3.2 of Unit 6 introduced the distribution of the number of values, x, above the median in a sample of 12 items. In this activity you are going to use Minitab to obtain these probabilities by doing the following.

- Open a new worksheet in Minitab. (If you have not just started a new session, create a new worksheet via **File** > **New**.)
- In the worksheet, create a column called `x` with the entries `0`, `1`, ..., `12`, to represent the values that `x` can take (**Calc** > **Make Patterned Data** > **Simple Set of Numbers**).
- Click on **Calc**, then **Probability Distributions**, and then click on **Binomial...**.
- In the **Binomial Distribution** dialogue box: select **Probability**; type `12` in the **Number of trials** field, as we have a sample of 12 observations; type `0.5` in the **Event probability** field, as 0.5 is the probability for the event that a single observation is above the population median; type `x` in the **Input column** field; type `prob` in the **Optional storage** field, to save the probabilities in a column labelled `prob`.

  The dialogue box should now look like Figure 26.

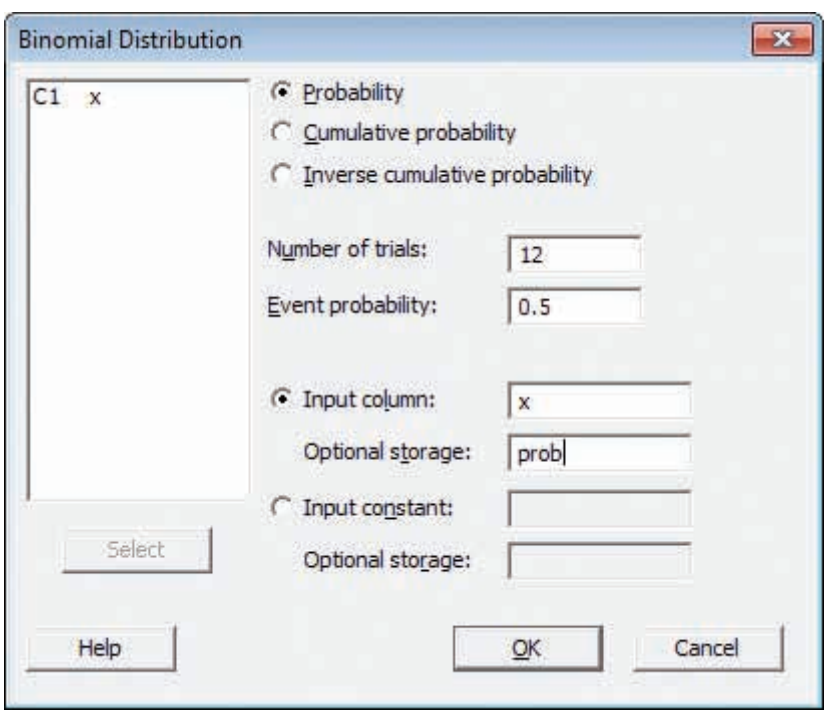


Figure 26 The **Binomial Distribution** dialogue box

- Click on **OK**.

A new column of numbers appears in the worksheet. These numbers are the values of $P(x = 0)$, $P(x = 1)$, ..., $P(x = 12)$. Check that when rounded to three decimal places, these match the values shown in Figure 5 of Unit 6 (Subsection 3.2). For convenience, this figure is reproduced below as Figure 27.

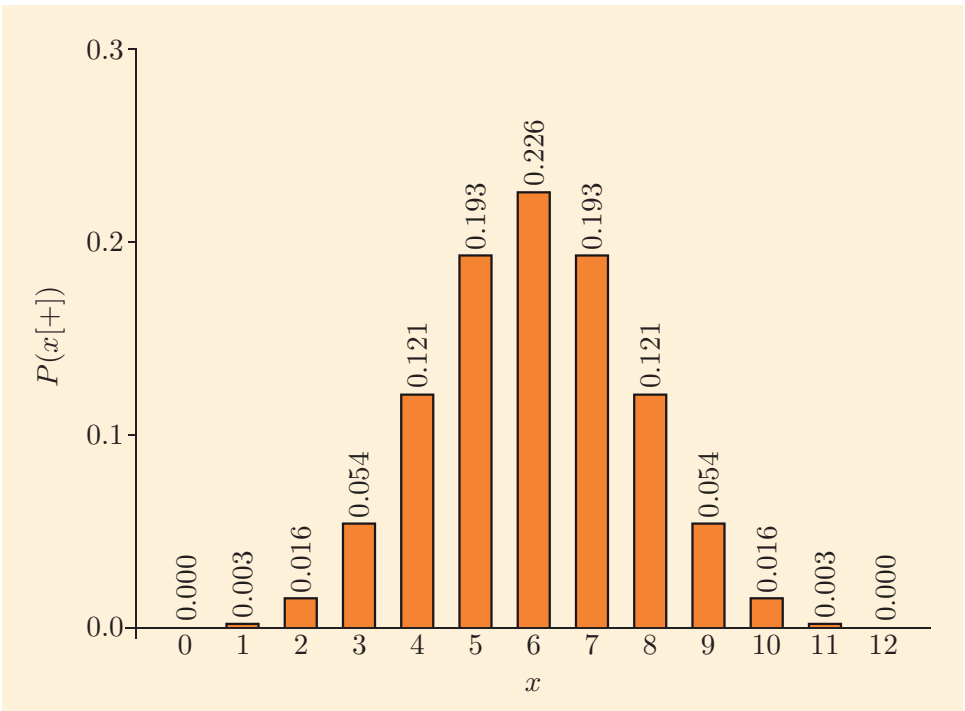


Figure 27 Probability distributions for a random sample of size 12

You will need the results in this worksheet for Computer activity 54. So, if possible, carry straight on to Computer activity 54.

Computer activity 54 *A bar chart of the probability distribution*

Figure 27 displays the probability distribution of the number of values above the median in a sample of 12. Use Minitab to produce a similar plot by doing the following.

- Click on **Graph** and select **Bar Chart...**.
- In the **Bar Charts** dialogue box, change the **Bars represent** field to `Values from a table`, using the drop-down menu. Leave **Simple** as the form of bar chart selected and click on **OK**.
- A dialogue box will open entitled **Bar Chart: Values from a table, One column of values, Simple**. Enter `prob` in the **Graph variables** field and enter `x` in the **Categorical variable** field. The dialogue box should now look like Figure 28.

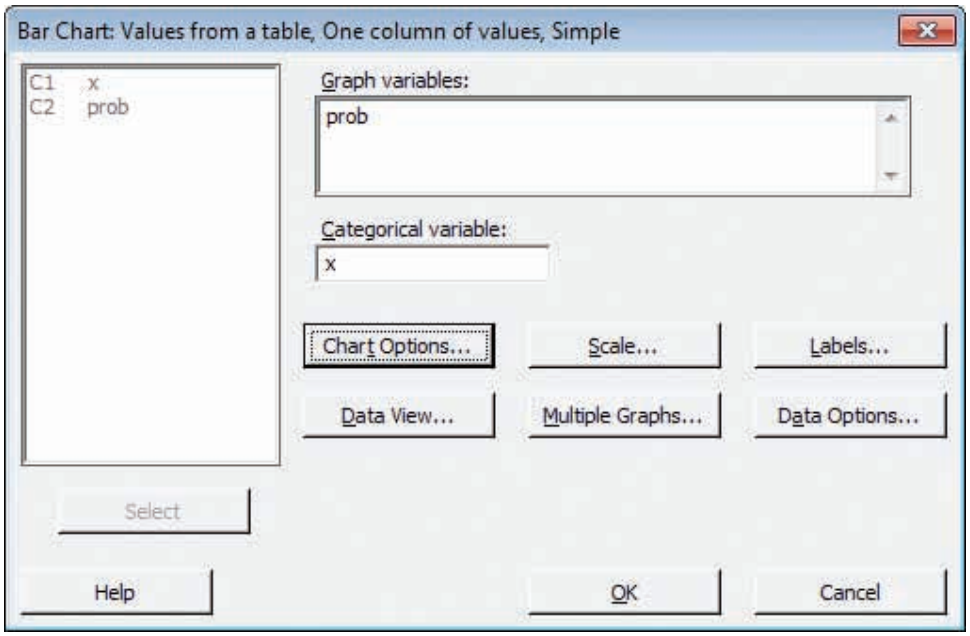


Figure 28 The **Bar Chart: Values from a table, One column of values, Simple** dialogue box

- Click on **OK**.

The resulting bar chart produced by Minitab is displayed in Figure 29. Notice this resembles Figure 27, apart from the axis labels.

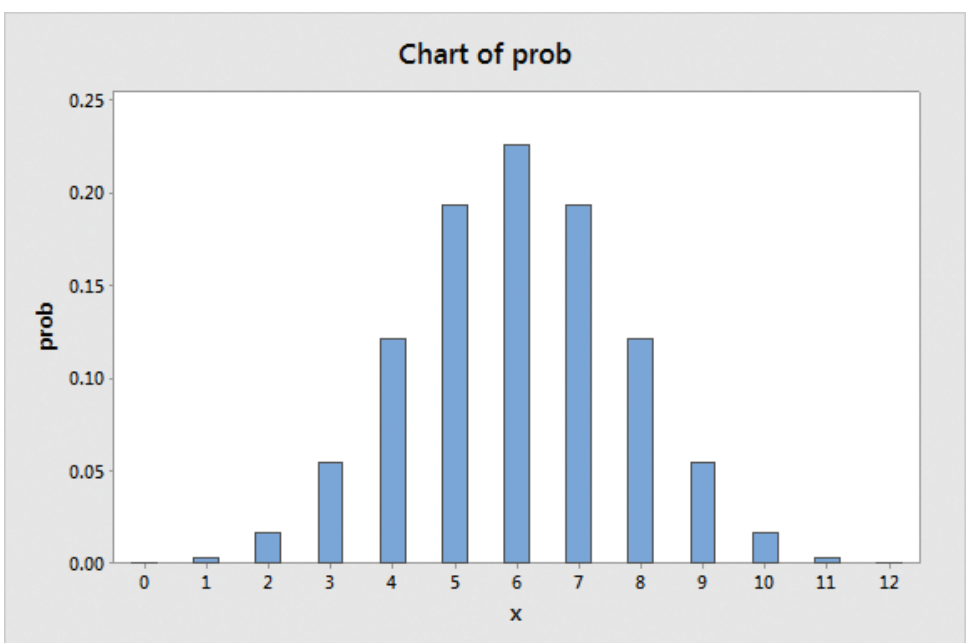


Figure 29 Bar chart of the probability distribution in Computer activity 53

Computer activity 55 *Cumulative probabilities*

In Computer activity 53, you used Minitab to calculate the probabilities $P(0[+])$, $P(1[+])$, ..., $P(12[+])$. However, as you saw in Unit 6, these probabilities are only important for the sign test because they can be used to calculate probabilities such as $P(0[+]) + P(1[+])$ and $P(0[+]) + P(1[+]) + P(2[+])$. These probabilities, known as **cumulative probabilities**, can be calculated directly in Minitab. To obtain these for the distribution used in Computer activity 53, do the following.

- Open a worksheet in Minitab, which has a column `x` with entries `0`, `1`, ..., `12`. It does not matter whether or not this worksheet also has a column called `prob` left over from Computer activity 53.
- Obtain the **Binomial Distribution** dialogue box (**Calc** > **Probability Distributions** > **Binomial**).
- In the **Binomial Distribution** dialogue box, this time select **Cumulative probability**. As before, enter `12` in the **Number of trials** field, `0.5` in the **Event probability** field and `x` in the **Input column** field. (You may find these entries already in the dialogue box if you are carrying straight on from Computer activity 53.) In the **Optional storage** field enter `cumprob`.
- Click on **OK**.

A new column of numbers called `cumprob` appears in the worksheet. These numbers are the values of $P(x \leq 0)$, $P(x \leq 1)$, ..., $P(x \leq 12)$.

(a) Use the numbers in the worksheet to write down $P(x \leq 4)$ and $P(x \leq 5)$.

(b) Look at the numbers going down the column `cumprob`. What do you notice? Why must it be this way?

(c) Write down $P(x \leq 12)$. Hence comment on how likely it is that there are 12 or fewer values above the median in a sample of 12.

The next activity will give you some practice with obtaining cumulative probabilities using Minitab.

Computer activity 56 *Probability distribution for a random sample of size 17*

Car batteries produced by one manufacturer have a median usable life of 60 months. Suppose 17 batteries are chosen at random and let x be the number of these batteries that have a usable life of more than 60 months.

(a) Use Minitab to obtain the cumulative probabilities $P(x \leq 0)$, $P(x \leq 1)$, ..., $P(x \leq 17)$.

(b) Use the cumulative probabilities to write down $P(x \leq 4)$ and $P(x \leq 5)$.

(c) In Table 8 of Unit 6 (Subsection 4.1), the critical value at the 5% significance level is given as 4 for a sample size of 17. Explain why your answers in part (b) of this activity are consistent with that.

6.2 The sign test

Hypothesis tests form an important feature of almost all statistics packages. It varies as to which hypothesis tests each performs; however, general purpose statistical packages (such as Minitab) all include a core of the common hypothesis tests. The purpose is to make it easy for the user to perform the tests and to obtain useful information. In later subsections of this Computer Book you will learn how to perform a number of hypothesis tests using Minitab. In Computer activity 57 you will learn how to perform the first of these tests: the sign test.

Computer activity 57 *The median weight of cats*

In this activity you are going to test the following hypothesis using the sign test:

The median weight of male cats equals 3.0 kg.

The data to be used for this test consist of the weights (in kg) of a random sample of 40 male cats. (Source: Chen, K.K., Bliss, C.I. and Robbins, E.B. (1942) 'The digitalis-like principles of *calotropis* compared with other cardiac substances', *Journal of Pharmacology and Experimental Therapeutics*, vol. 74, pp. 223–34.) These weights are given in the Minitab worksheet **catweights.mtw**.

- Open the worksheet **catweights.mtw** in Minitab and make sure it is the active worksheet. Notice that the weights of the cats are given in the column `weight`.
- Click on **Stat**, then **Nonparametrics**, and then **1-Sample Sign...**, to open the **1-Sample Sign** dialogue box.
- In the **1-Sample Sign** dialogue box, copy `weight` in the **Variables** field. Select **Test median** and type `3.0` in the associated field, as that is the value hypothesised for the median. Make sure that `not equal` is given in the **Alternative** field. The completed dialogue box should look like Figure 30.

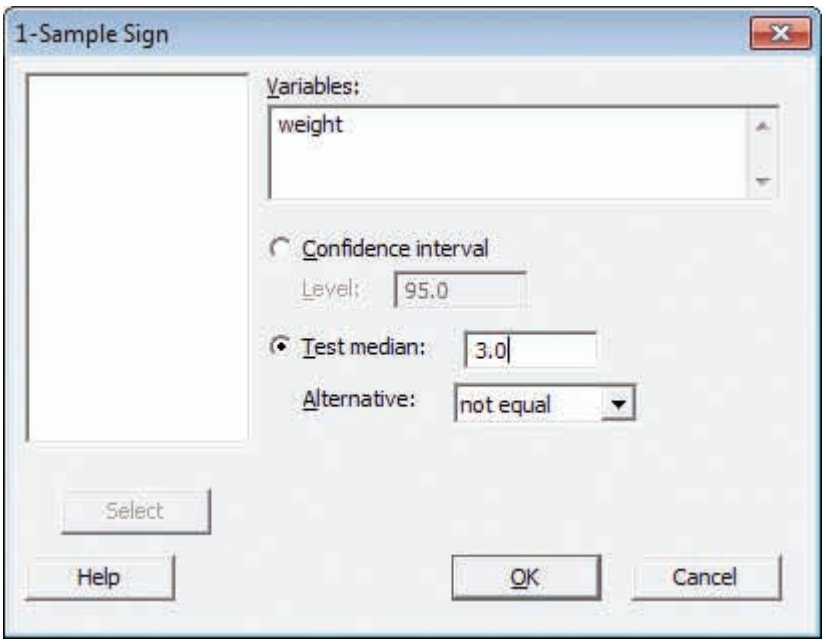


Figure 30 The **1-Sample Sign** dialogue box

- Click on **OK**.

The output from the sign test produced by Minitab includes the following.

```
Sign test of median =  3.000 versus ≠ 3.000

          N  Below  Equal  Above       P  Median
weight   40     27      4      9  0.0039   2.600
```

The first line states that you have used the sign test to test whether the population median equals 3.000 (as you specified). This provides a check of the hypothesis that Minitab has used when carrying out the calculations for the sign test. The extra decimal places that Minitab gives for the hypothesised population median indicate the extra accuracy it uses when determining whether a value is a tie. The bottom two lines give the following results of the analysis.

- The value for `N` indicates there were 40 cats.
- The value for `Below` indicates that 27 of the cats had a weight of less than 3.0 (kg).
- The value for `Equal` indicates that 4 of the cats had a weight of exactly 3.0 (kg).
- The value for `Above` indicates that 9 of the cats had a weight above 3.0 (kg).
- The value for `Median` indicates that the median weight of cats in the sample was 2.600 (kg).

- The value for `P` gives the most important figure in the output – the p-value. Here, the p-value equals 0.0039.

Recall that in Subsection 5.1 of Unit 6 the following guidelines for the interpretation of p-values were given.

Table 1 Interpretation of p-values

p-value	Rough interpretation
$p > 0.10$	Little evidence against the hypothesis
$0.10 \geq p > 0.05$	Weak evidence against the hypothesis
$0.05 \geq p > 0.01$	Moderate evidence against the hypothesis
$0.01 \geq p > 0.001$	Strong evidence against the hypothesis
$0.001 \geq p$	Very strong evidence against the hypothesis

A p-value of 0.0039 lies between 0.01 and 0.001, so from Table 1, there is strong evidence against the hypothesis – that is, strong evidence against the hypothesis that the population median weight of male cats is 3.0 kg. We therefore conclude that there is strong evidence that the median weight of male cats in the population is not 3.0 kg. In fact, as the sample median is 2.6 kg, the data suggest that the population median weight of male cats is less than 3.0 kg.

The next activity will give you some practice at doing the sign test.

Computer activity 58 *Testing claims about petrol consumption*

In Exercise 6, at the end of Section 4 of Unit 6, data on the petrol consumption of a Honda Civic 1.4i were given and a car dealer's claim that the car managed 37 miles per gallon was tested. In this activity you will test other claims that might have been made about the car's petrol consumption.

The Minitab worksheet **petrol3.mtw** contains the data on the petrol consumption of a Honda Civic 1.4i. In this worksheet, the variable `mpg` contains the miles per gallon achieved between a series of refuelling stops and is the basis for the stemplot given in Exercise 6 of Unit 6.

(a) Suppose a car dealer claimed that the median petrol consumption of a Honda Civic 1.4i is 36 miles per gallon. Use Minitab to perform a sign test to test whether the car dealer's claim is plausible. Give the p-value from the test and say what may be concluded from the test in simple terms.

(b) Suppose the car dealer had claimed that the car would give 37 miles per gallon. Repeat part (a) for this claim.

Summary of Chapter 6

In this chapter, you have learned how to use Minitab to calculate probabilities connected with a sign test. You have seen how to work out the probability that a particular number of observations are greater than the population median, and so you can work out the p-value for yourself. You have also learned how to obtain the p-value directly using Minitab.

7 The normal distribution

This chapter, which is associated with Unit 7, focuses on the normal distribution. In Subsection 7.1 you will use an interactive computer resource to compare normal distributions with different means and standard deviations. The transformation of normal distributions to the standard normal distribution is the focus of Subsection 7.2. Then in Subsection 7.3 you will learn how use Minitab to do the one-sample z-test (a test that makes use of the normal distribution).

7.1 Exploring the normal distribution

In this subsection you will explore the location and spread of a normal distribution, as measured by the mean μ and standard deviation σ.

Computer activity 59 *Effects on the normal distribution of changing the mean*

Open the interactive computer resource 'Normal distribution' on the M140 website.

This resource shows the normal distribution, and allows you to change the mean, μ, and the standard deviation, σ. The initial setting is $\mu = 0$, $\sigma = 1$; the corresponding normal distribution is the same as that shown in Figure 11(a) of Unit 7 (Subsection 3.1). For convenience, Figure 11 is reproduced as Figure 31.

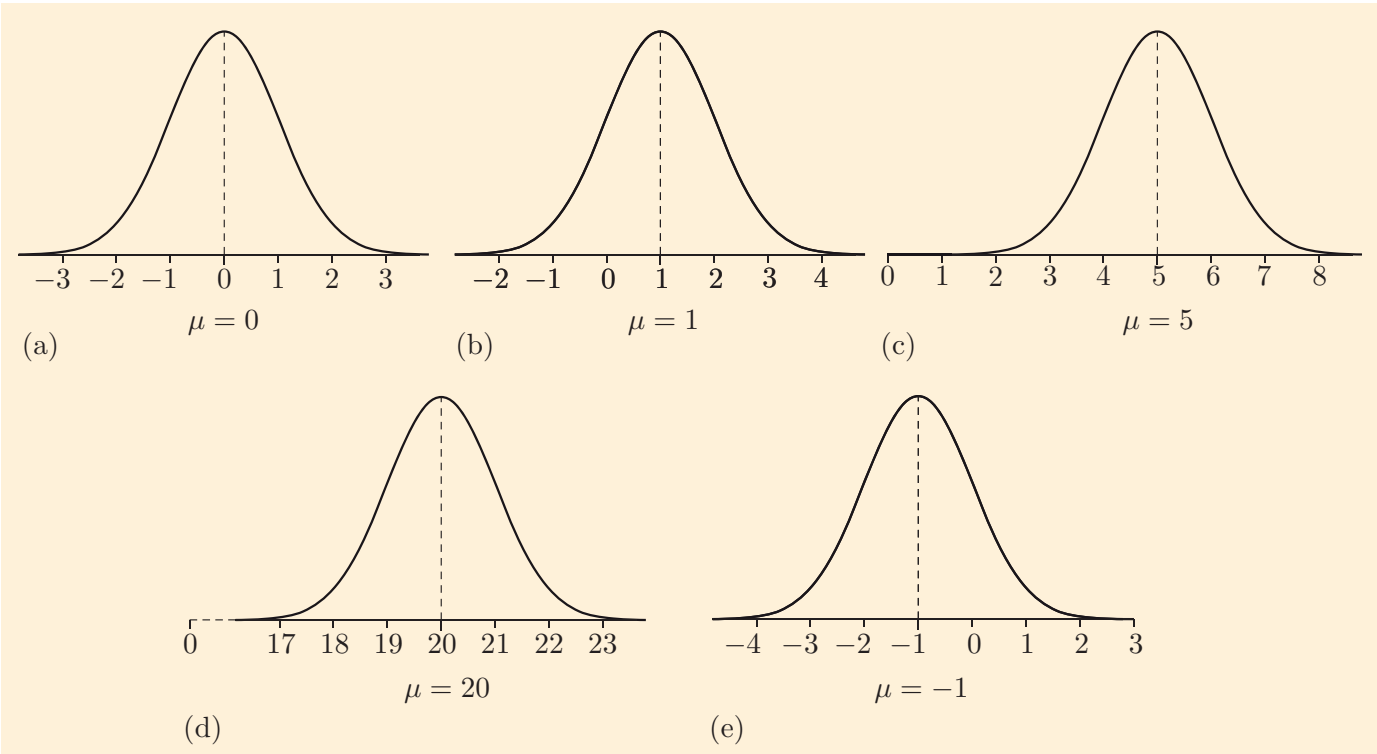


Figure 31 Normal distributions with different locations

(a) Change the value of μ from 0 to 1, leaving the value for σ at its initial setting. Which part of Figure 31 is now displayed?

(b) Next, change the value of μ to -1. Which part of Figure 31 is now displayed?

(c) By varying the value of μ over the range allowed, describe what happens to the normal distribution in the resource relative to its starting position.

Computer activity 60 *Effects on the normal distribution of changing the standard deviation*

In Computer activity 59, you explored the effect of changing the mean, μ, of a normal distribution. In this activity, you will explore the effect of changing the standard deviation, σ.

- Open the interactive computer resource 'Normal distribution' if it is not already open.
- Click on 'Reset'.

For the initial setting $\mu = 0$, $\sigma = 1$, the corresponding normal distribution is the same as that shown in Figure 12(b) of Unit 7 (Subsection 3.1). For convenience, Figure 12 is reproduced as Figure 32.

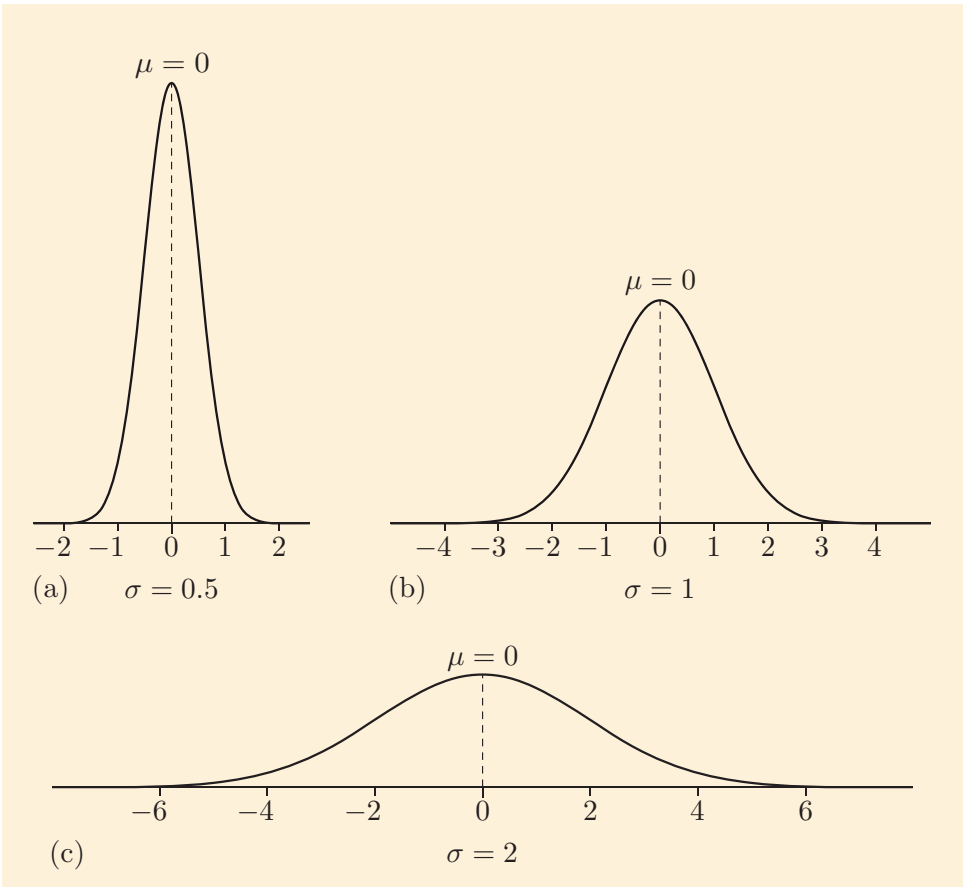


Figure 32 Normal distributions with different spreads

(a) Change the value of σ from 1 to 2. Which part of Figure 32 is now displayed?

(b) Next, change the value of σ to 0.5. Which part of Figure 32 is now displayed?

(c) By varying the value of σ over the range allowed, describe what happens to the normal distribution relative to its starting position.

Computer activity 61 *Effects on the normal distribution of changing both the mean and the standard deviation*

So far in this subsection you have used the interactive computer resource 'Normal distribution' to explore the effect of changing either the mean, μ, or the standard deviation, σ. In this activity, you will explore the effect of changing the mean and standard deviation at the same time.

Make sure that you have the interactive computer resource 'Normal distribution' open and that the initial values are set at $\mu = 0$ and $\sigma = 1$.

(a) Change the value of μ from 0 to 1, and value of σ from 1 to 2. Describe what happens to the position and shape of the normal distribution relative to its starting position.

(b) Two of the following three statements are incorrect. By experimenting with the values of μ and σ in the resource, identify which statement is correct and what is wrong with the other two statements.

A. 'When $\mu > 0$ and $\sigma < 1$, the normal distribution is moved to the left and is less spread out relative to the normal distribution with $\mu = 0$, $\sigma = 1$.'

B. 'When $\mu > 0$ and $\sigma = 1$, the normal distribution is moved to the right and has the same spread relative to the normal distribution with $\mu = 0$, $\sigma = 1$.'

C. 'When $\mu < 0$ and $\sigma > 1$, the normal distribution is moved to the left and is less spread out relative to the normal distribution with $\mu = 0$, $\sigma = 1$.'

7.2 Transforming normal distributions

In Subsection 3.3 of Unit 7, you learned that a variable that has a normal distribution with mean μ and standard deviation σ can be transformed to a variable z that has the standard normal distribution (that is, a normal distribution with mean 0 and standard deviation 1). In particular, you learned that when the original variable is denoted by x, the transformation has the form

$$z = \frac{x - \mu}{\sigma}.$$

In this subsection, you will use an interactive computer resource to verify that this transformation is the right one to use.

Computer activity 62 *Transforming a normal distribution to the standard normal distribution*

Open the interactive computer resource 'Transforming normal distributions'. This resource allows you to display and move a normal distribution so that it matches the standard normal distribution.

(a) The distribution of a particular variable is normal and has mean $\mu = 1$ and standard deviation $\sigma = 2$. Enter the relevant values for μ and σ so that the corresponding normal distribution curve is displayed. Note that the standard normal distribution will remain showing in the background.

Changing the value of a shifts the whole distribution left or right. (That is, it changes the *location* of the distribution but leaves the *spread* unchanged.) It corresponds to the transformation $v = x - a$. What value of a moves this normal distribution so that it has the same mode (0) as the standard normal distribution?

(b) The initial curve from part (a) is now a new normal distribution with mean 0 and standard deviation 2 (unchanged).

Changing the value of b rescales this distribution by moving the peak of the curve up or down. (That is, it changes the *spread* of the distribution but leaves the location unchanged.) It corresponds to the transformation $z = v/b$. What value of b moves this new normal distribution so that is has the same spread as the standard normal distribution, and hence matches the standard normal distribution?

Before moving on to Computer activity 63, let's revise what the results of Computer activity 62 mean in a little more detail.

1. We started with a normal distribution of a variable that it will now be convenient to call x, with mean 1 and standard deviation 2.
2. By subtracting 1 from each value of x we obtained the distribution of a new variable v, where $v = x - 1$. The distribution of v was also normal but with mean 0 and standard deviation 2.
3. Then we divided each value of v by 2 to get the variable z, where $z = v/2$. This variable z has the standard normal distribution.
4. The variables x and z are therefore related by the formula $z = (x - 1)/2$.

Computer activity 63 *Transforming more normal distributions*

Open the interactive computer resource 'Transforming normal distributions', if it is not already open. Click on 'Reset'.

(a) The distribution of another variable is normal, and has mean $\mu = 2.5$ and standard deviation $\sigma = 0.5$. Repeat Computer activity 62 for this normal distribution, and report the values of a and b which transform it into the standard normal distribution.

(b) The distribution of yet another variable is normal, and has mean $\mu = -1.5$ and standard deviation $\sigma = 1.5$. Repeat Computer activity 62 for this normal distribution, and report the values of a and b which transform it into the standard normal distribution.

(c) Using your experiences above, does $z = (x - \mu)/\sigma$ appear to be the right formula to transform a normal distribution with mean μ and standard deviation σ to the standard normal distribution?

7.3 The one-sample z-test using Minitab

In Section 5 of Unit 7, you investigated the one-sample z-test and completed the calculations associated with the one-sample z-test 'by hand' (that is, using a calculator). Similarly to the sign test, it is possible to use Minitab to do one-sample z-tests. (Minitab does not provide a facility for performing two-sample z-tests.)

The use of Minitab to do one-sample z-tests will be introduced by re-running the first hypothesis test performed in Subsection 5.2 of Unit 7.

Computer activity 64 *Testing the average reading ability of seven-year-olds*

Based on data from the British Cohort Study (BCS), the first hypothesis test in Subsection 5.2 of Unit 7 concerned testing

H_0: $\mu = 96$

H_1: $\mu \neq 96$,

where μ is the population mean of the British Ability Scales reading scores of all British 7-year-old children in 2004–2005. The data from the BCS concerning 7-year-old children are summarised in Table 2, which is a copy of Table 3 in Unit 7 (Subsection 5.2).

Table 2 Summary statistics for the data on reading scores of 7-year-old children

Sample size	Sample mean	Sample standard deviation
396	111.28	26.668

(These data are copyright and owned by the Economic and Social Data Service.)

Notice that the data are given in a summarised form. This means that a worksheet containing the data does not need to be opened in Minitab first. Instead, this summary information will be entered directly into the **One-Sample Z for the Mean** dialogue box. So in Minitab do the following.

- Click on **Stat**, choose **Basic Statistics**, and then choose **1-Sample Z...**. The **One-Sample Z for the Mean** dialogue box will appear.
- You are going to be entering the data in a summarised form. So, in the dialogue box, select `Summarized data` from the top drop-down list. Notice that fields for the **Sample size** and **Sample mean** now appear in the dialogue box.
- Type `396` in the **Sample size** field and `111.28` in the **Sample Mean** field.
- Type `26.668` in the **Known standard deviation** field.
- Minitab also needs to know details about the null hypothesis, so select **Perform hypothesis test**. The **Hypothesized mean** field becomes active. From the hypotheses given before Table 2, type `96` in the **Hypothesized mean** field.

The completed dialogue box should be as in Figure 33.

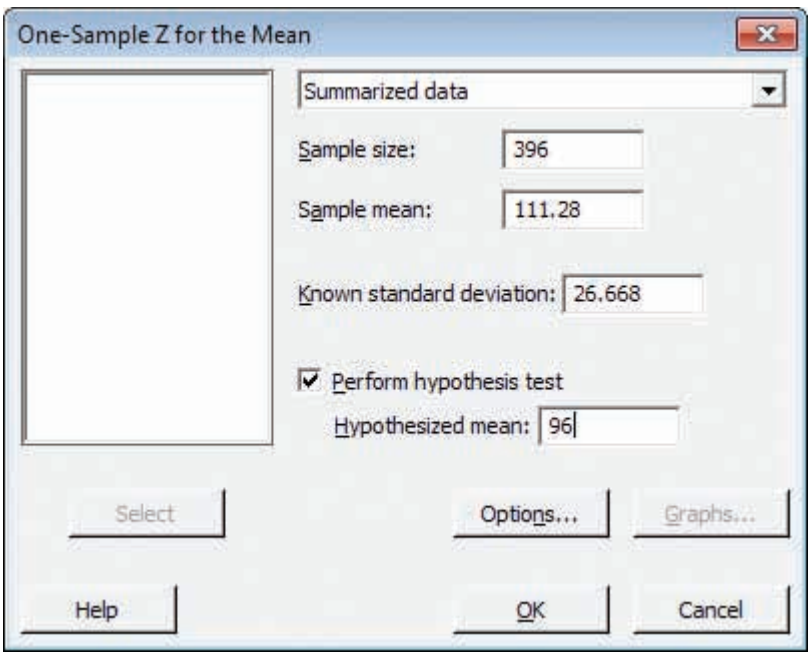


Figure 33 The **One-Sample Z for the Mean** dialogue box

- Click on **OK**.

Minitab immediately performs the test. The resulting Minitab output includes the following:

```
Test of μ = 96 vs ≠ 96
The assumed standard deviation = 26.668


  N    Mean  SE Mean        95% CI            Z      P
396  111.28     1.34  (108.65, 113.91)   11.40  0.000
```

What do we have here? Well, first the null and alternative hypotheses are quoted back at us; the two-sided nature of the test is confirmed by the alternative hypothesis, $H_1 : \mu \neq 96$. The standard deviation is quoted next. Then, at the beginning of the main line of results, we have the sample size (called `N`) and the (sample) mean. (These provide useful checks that you typed in the data and hypothesised values correctly.) The subsequent entries in the main line of results are:

- `SE Mean`, which is what we called ESE – the estimated standard error (of the mean) – in Subsection 5.2 of Unit 7
- `95% CI`, which is a confidence interval for μ – this is very useful in general, but is not covered in Unit 7
- `Z`, which is what we called z in Subsection 5.2 of Unit 7 – its value, `11.40`, coincides with that calculated by hand in Example 7 of Unit 7 (Subsection 5.2)

Minitab is not omniscient! It calls the population mean `mu`, just as we have; but that is because it always calls the population mean `mu` whatever we call it.

- `P`, which has a value of `0.000` and is the final entry – just as with the output from the sign test, `P` gives the p-value for the test.

Here the value of p given by Minitab for the BCS hypothesis test considered above is $p = 0.000$. This does not mean that $p = 0$; Minitab gives its results rounded to three decimal places. According to Table 1 (Subsection 6.2), it does, however, correspond to very strong evidence against the null hypothesis: $0.001 \geq p$. In other words, there is very strong evidence that the mean reading ability of 7-year-old children in 2004–2005 is not 96. In fact, the data suggest that the mean reading ability is more than 96.

Note that the result of the BCS hypothesis test given in the solution to Example 7 of Unit 7 (Subsection 5.2) was to reject the null hypothesis at the 1% level; this corresponds to $0.01 \geq p$ above. So by using Minitab, you have been able to determine that there is more evidence against the null hypothesis than you were able to find when doing the test by hand.

The following two activities provide more practice in carrying out the one-sample z-test using Minitab.

Computer activity 65 *Is the mean weekly wage of male leisure and sports managers equal to the overall mean weekly wage for male employees?*

A random sample of 230 male leisure and sports managers had a mean wage of £587 per week in 2011 with a standard deviation of £208. The overall mean weekly wage for male employees in 2011 was £598. (Source: *Annual Survey of Hours and Earnings*, 2011.)

Use Minitab to investigate whether the mean weekly wage of male leisure and sports managers differed from the overall mean weekly wage for male employees in 2011. Comment on your result.

Computer activity 66 *Is the mean weight of glass in milk bottles constant?*

Glass milk bottles are formed from molten lumps ('gobs') of glass that can be weighed as they fall into moulds. It is important for one manufacturer that the mean weight of glass per bottle is maintained at 255 grams; too low a mean results in too many fragile bottles, and too high a mean leads to an excessive consumption of glass.

The standard deviation of the weight per bottle is known to be 2.5 grams. Sometimes the mean weight of glass per bottle delivered by the machine changes slightly, though the standard deviation is most unlikely to vary. Accordingly a sample of bottles is taken and weighed periodically. On one occasion a sample of 27 bottles had a mean weight of 256.19 grams. Is there any reason to adjust the machine?

Summary of Chapter 7

This chapter has focused on the normal distribution. You have seen that the mode of the distribution is the same as the mean, μ, and that as the standard deviation increases, the distribution becomes flatter (but still has the same general shape). You have also verified, using some specific examples, that the formula $z = (x - \mu)/\sigma$ transforms a variable x that has a normal distribution with mean μ and standard deviation σ, to a variable z that has a standard normal distribution.

You have also learned how to do one-sample z-tests in Minitab and interpreted the results.

8 The χ^2 test for contingency tables

In this chapter, which is associated with Unit 8, you will learn how to perform a χ^2 test using Minitab. In Subsection 8.1 you will learn how to do this using a contingency table that has been stored in a Minitab worksheet. Then, in Subsection 8.2, you will learn how to enter your own contingency tables in Minitab.

8.1 Doing the χ^2 test in Minitab

In this subsection you will learn how to use Minitab to perform a χ^2 test using data that have already been stored in a Minitab worksheet. The contingency table which will first be analysed in this subsection is based on data from the Clackmannanshire study which was discussed in Example 2 of Unit 8 (Subsection 2.1). These data relate to reading ability in the three teaching method groups: analytic phonics, analytic phonics + phonological awareness (PA), and synthetic phonics. This contingency table (*without* the marginal totals) is reproduced below as Table 3.

Table 3 Results from the baseline test

	Reading age as compared to chronological age	
	Not higher	Higher
Analytic phonics	69	39
Analytic phonics + PA	43	35
Synthetic phonics	76	41

Computer activity 67 *Looking at contingency table data in Minitab*

The data in Table 3 have been entered into the Minitab worksheet **baselineR.mtw**. Open this worksheet in Minitab now.

Look at how the data have been laid out in the worksheet. How does this correspond to Table 3? Are the marginal totals included in the worksheet?

If possible keep this worksheet open in Minitab as you will need it for the next activity.

Computer activity 68 *Obtaining the χ^2 test statistic*

The solution to Activity 14 in Unit 8 (Subsection 4.1) gave the following hypotheses, which are relevant to Table 3.

H_0: Teaching method for children starting at primary school and reading ability at the baseline test are independent.

H_1: Teaching method for children starting at primary school and reading ability at the baseline test are not independent.

In Unit 8 you tested these hypotheses using the χ^2 test for contingency tables. In this activity you will use Minitab to obtain the χ^2 statistic for the data in Table 3, and interpret the corresponding output.

- Make sure that the worksheet **baselineR.mtw** is the active worksheet in Minitab.
- Click on **Stat**, then on **Tables**, and select **Chi-Square Test for Association...**.
- The **Chi-Square Test for Association...** dialogue box will appear. From the top drop-down list select `Summarized data in a two-way table`.

  Copy the columns containing the data on which you want to do the χ^2 test into the **Columns containing the table** field. These are `nothigher` and `higher`.
- In the **Rows** field add `'Teaching group'`. This will ensure that the rows will be labelled correctly in the output.

  The completed dialogue box should be as in Figure 34.

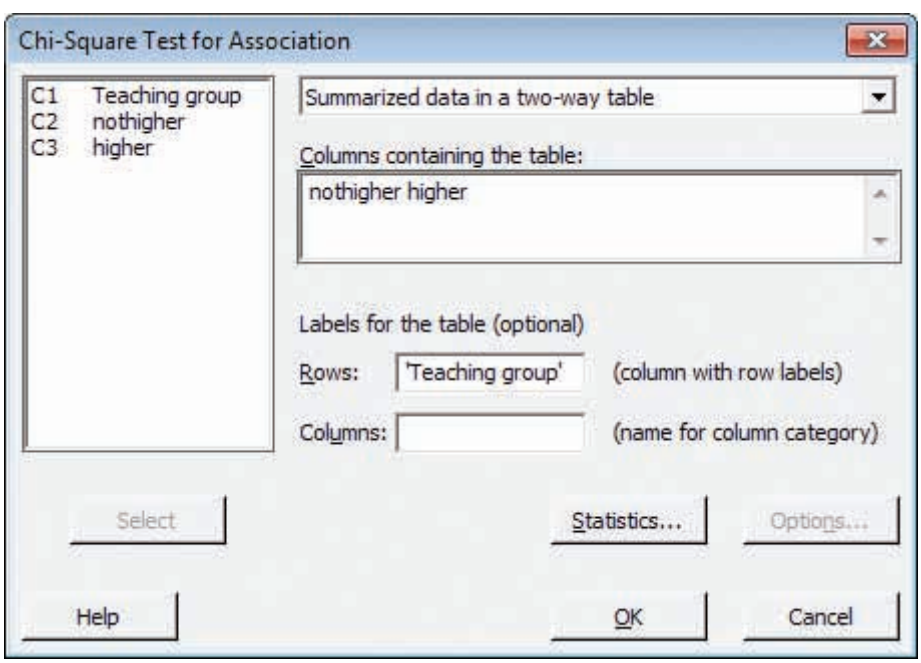


Figure 34 The **Chi-Square Test for Association** dialogue box

- Now click on **OK**.

The following text, containing details of the χ^2 test results, will then appear in the Session window.

```
Rows: Teaching group   Columns: Worksheet columns

                          nothigher  higher  All

Analytic phonics                 69      39  108
                              67.01   40.99

Analytic phonics + PA            43      35   78
                              48.40   29.60

Synthetic phonics                76      41  117
                              72.59   44.41

All                             188     115  303

Cell Contents:      Count
                    Expected count


Pearson Chi-Square = 2.162, DF = 2, P-Value = 0.339
Likelihood Ratio Chi-Square = 2.136, DF = 2, P-Value = 0.344
```

The display shows three horizontal bands. These correspond to three categories of `Teaching group`: `Analytic phonics`, `Analytic phonics + PA`, and `Synthetic phonics`. Within each band, the first row contains the counts (taken from the worksheet). Under each count is the corresponding Expected value. (You can check that these values match those obtained in Activity 16 of Unit 8 (Subsection 4.2).) The display also includes row and column totals. Details of two types of χ^2 test statistic are given at the bottom of the display. The version we have been using in M140 corresponds to the Pearson Chi-Square line. Thus, `Pearson Chi-Square = 2.162` indicates that the χ^2 test statistic is 2.162; `DF = 2` tells you that the degrees of freedom of this contingency table are 2; and `P-Value = 0.339` tells you that the p-value for the test is 0.339.

In Unit 8, the χ^2 test statistic was compared to CV5 and CV1, the 5% and 1% critical values of the appropriate χ^2 distribution. Minitab instead gives you the p-value, or significance probability, just as it did for the sign test and for the z-test. The (rough) interpretation of p-values was given in Table 1 (Subsection 6.2). Notice that in Table 1 the interpretation of the test result is in terms of *the strength of evidence* against the null hypothesis, whereas the approach described in Unit 8 provides an interpretation in terms of *the decision to reject or not to reject* the null hypothesis. These interpretations are essentially equivalent.

The advantage of using CV5 and CV1 is that you can then do a χ^2 test without a computer.

In this case, the p-value is more than 0.10 so there is little evidence against the null hypothesis that the teaching method for children starting at primary school and reading ability at the baseline test are independent.

The next two activities will give you some practice at doing the χ^2 test in Minitab and interpreting its results.

Computer activity 69 *Reading ability at the first follow-up test*

The Minitab worksheet **firstR.mtw** contains the data on reading abilities at the first follow-up test in the Clackmannanshire study. These data were presented in Table 3 of Unit 8 (Subsection 2.1).

These data allow the following hypotheses to be tested using the χ^2 test:

H_0: Teaching method for children starting at primary school and reading ability at the first follow-up test are independent.

H_1: Teaching method for children starting at primary school and reading ability at the first follow-up test are not independent.

(a) Perform the χ^2 test using Minitab. Identify the Expected value for the `higher` category in the `Analytic phonics` group, the χ^2 contribution from the `nothigher` category in the `Synthetic phonics` group, and the χ^2 test statistic.

(b) Interpret the results of the χ^2 test, in terms of strength of evidence against the null hypothesis and in terms of a decision to reject or not reject the null hypothesis at the 5% and 1% significance levels. What do you conclude?

Next, in Computer activity 70, you are asked to perform a χ^2 analysis of a larger contingency table. In the last part, you are asked to describe any departure from independence you might observe. Minitab does not print residuals, but their sign can easily be deduced by comparing the count and Expected value for individual cells of interest.

Computer activity 70 *Reading ability at baseline and first follow-up tests*

The Minitab worksheet **basefirstR.mtw** contains the data on reading abilities at the baseline and first follow-up tests in the Clackmannanshire study. These data were presented in Table 6 of Unit 8 (Subsection 2.2). Open this worksheet now.

The data are now in four columns, `lowlow` to `highhigh`. For example, the column called `highlow` contains the numbers of children in each group who scored Higher in the baseline test and Not higher in the first follow-up test.

(a) Perform the χ^2 test using Minitab. Check that the degrees of freedom given by Minitab agree with the value given by the formula $(r-1)\times(c-1)$.

(b) Explain why the results given by Minitab suggest that there is an association between the reading ability at baseline and first follow-up tests and the teaching method used.

8.2 Entering contingency tables into Minitab

In this subsection you will learn how to enter contingency tables into a Minitab worksheet prior to analysis. This is obviously needed to analyse new data, but also, as you shall see, when you need to combine rows or columns because one or more Expected values is less than 5.

Computer activity 71 *Spelling and reading ability at the second follow-up tests*

Table 4 contains data on the spelling and reading abilities of children at the second follow-up tests, by teaching method. You met similar data (for the first follow-up test) in Example 9 of Unit 8 (Subsection 3.2). For example, the category Low/High here represents children whose spelling age is lower but whose reading age is higher than chronological age.

Table 4 Spelling and reading ability at second follow-up tests

	Spelling/Reading age at second follow-up tests				
	Low/Low	Low/High	High/Low	High/High	Total
Analytic phonics	7	3	11	74	95
Analytic phonics + PA	6	1	5	54	66
Synthetic phonics	8	2	7	88	105
Total	21	6	23	216	266

Enter this contingency table into a Minitab worksheet by doing the following.

- Open up a new worksheet in Minitab (**File** > **New**).
- Enter the row categories `Analytic phonics`, `Analytic phonics + PA` and `Synthetic phonics` in rows 1 to 3 of column `C1`, which will change to `C1-T`. You may need to resize the first column so all the categories can be read; to do this, place the mouse pointer on the right column boundary in the `C1-T` cell, and drag it to the right.
- In the grey box just under `C1-T` enter the variable name `Teaching group`.
- In the grey cells below `C2` to `C5` enter the column categories `lowlow`, `lowhigh`, `highlow`, `highhigh`.

  This completes the labelling of the contingency table. (No marginal totals will be entered, so no labelling is needed for these.)

  The labelling of the table is not actually required to do the χ^2 test. However, it is good practice to label data, in order to keep a record of which variables and categories are used.
- The next stage is to enter the counts. There are 3 rows and 4 columns of data to enter, so 12 counts. (The marginal totals should not be entered into the worksheet.) Enter the data now.

You should end up with the worksheet looking as in Figure 35. (A copy of the correct worksheet is given in the file **secondSR.mtw**.)

↓	C1-T	C2	C3	C4	C5	C6
	Teaching group	lowlow	lowhigh	highlow	highhigh	
1	Analytic phonics	7	3	11	74	
2	Analytic phonics + PA	6	1	5	54	
3	Synthetic phonics	8	2	7	88	
4						

Figure 35 A Minitab worksheet containing the data in Table 4

Now perform the χ^2 test on this contingency table to investigate the hypotheses:

H_0: Teaching method originally allocated and spelling/reading ability at the second follow-up tests are independent.

H_1: Teaching method originally allocated and spelling/reading ability at the second follow-up tests are not independent.

The now-familiar output will appear in the Session window, but with an added line to act as a warning at the end:

```
* NOTE * 3 cells with expected counts less than 5
```

The χ^2 test may be invalid, as some Expected values are less than 5. Can you identify the problem cells, and suggest a way round this problem?

In the solution to Computer activity 71 it was suggested that two columns of Table 4 should be combined to get round the problem of low Expected values. This is what you will do in Computer activity 72.

Computer activity 72 *Combining columns*

In this activity you will combine the `lowhigh` and `highlow` categories of Table 4 into a single `mixed` category, and analyse the new contingency table.

(a) By hand, create a contingency table where the categories have been combined. Enter this new contingency table into a Minitab worksheet and save it.

(b) Perform the χ^2 test on the new contingency table from part (a), and check that the Expected values are adequate. Interpret your findings.

Summary of Chapter 8

In this chapter, you have learned how to perform the χ^2 test in Minitab for contingency tables in Minitab worksheets. The interpretation of the results in terms of p-values was described. You have also learned how to input your own contingency tables into Minitab worksheets.

9 Correlation and interval estimates

This chapter, which is associated with Unit 9, focuses on correlation and interval estimates. In Subsection 9.1, you will learn how to calculate correlation coefficients using Minitab. The other three subsections in this chapter deal with interval estimates. In Subsection 9.2, you will learn how to use Minitab to obtain confidence intervals for a population mean based on the one-sample z-test. In Subsection 9.3, you will use a couple of interactive computer resources to explore the properties of confidence intervals. Finally, in Subsection 9.4, you will learn how to use Minitab to obtain interval estimates that are relevant for fitted lines: confidence intervals for the mean response, and prediction intervals.

9.1 Correlation coefficients

In this subsection you will learn how to obtain correlation coefficients using Minitab.

Computer activity 73 *Obtaining a correlation coefficient*

In Subsection 2.2 of Unit 9 you calculated a correlation coefficient for some data for eight constituencies in Wales: the percentage of children achieving Level 4 in specified subjects at Key Stage 2 in 2006 and the equivalent of at least five grade A* to C GCSEs including Mathematics and English or Welsh first language at Key Stage 4 in 2011. The data used in Examples 3 to 5 of Unit 9 are given in the worksheet **Welsh KS2 and KS4 results.mtw**. Open this worksheet in Minitab now.

Notice that the Key Stage 2 results are given in the variable `KS2` and the Key Stage 4 results are given in the variable `KS4`.

- Click on **Stat**, then **Basic Statistics**, and select **Correlation ...**. The **Correlation** dialogue box will now open.
- The aim is to calculate the correlation between the Key Stage 2 results and the Key Stage 4 results. So add both `KS2` and `KS4` to the **Variables** field. The completed dialogue box is as in Figure 36.

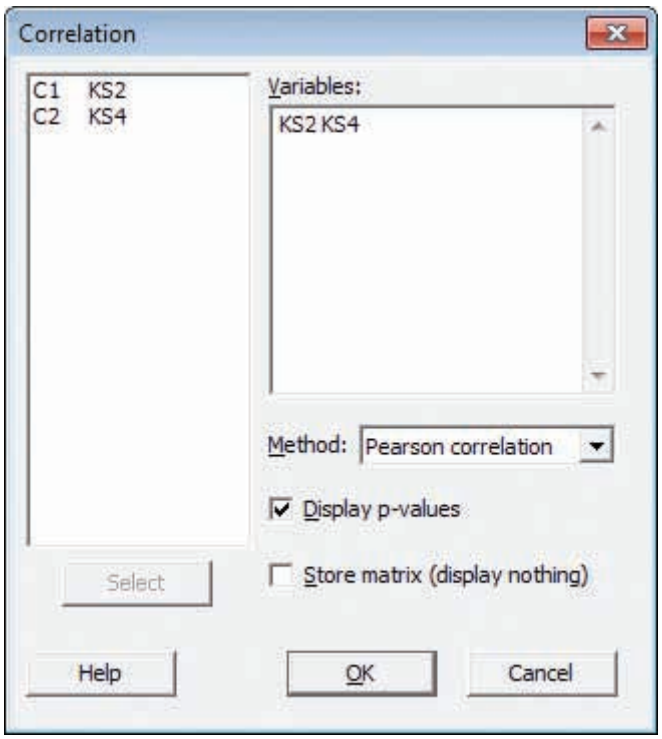


Figure 36 The **Correlation** dialogue box

- Click on **OK**.

The output in the Session window includes the line:

```
Pearson correlation of KS2 and KS4 = 0.639
```

So the value of the correlation coefficient given by Minitab is 0.639.

Computer activity 74 *Exploring relationships using correlation*

The data published for English secondary schools in 2011 includes information about pass rates at Key Stage 4 in different subject areas. The pass rates for 100 non-selective schools are given in the file **ks4subjects.mtw**. Open this worksheet in Minitab now.

In this worksheet the variables refer to the pass rates (as percentages) in the following subjects: `english` – English, `maths` – Mathematics, `science` – Science, `humanities` – History or Geography, and `language` – Languages. Note that an asterisk in a cell of a worksheet means that the corresponding data value is missing.

(a) Using Minitab, calculate the correlation coefficient between the pass rates in English and Mathematics. Interpret this correlation coefficient.

(b) Again using Minitab, calculate the correlation coefficients between the following pairs of subjects:

- English and Science
- English and History or Geography
- English and Languages.

Hence with which subject does the pass rate in English have the strongest relationship, and with which subject does the pass rate in English have the weakest relationship?

(c) Use Minitab to calculate the correlation coefficient between the pass rate for English and itself. (That is, obtain the correlation coefficient between `english` and `english`.)

9.2 Obtaining confidence intervals based on a one-sample z-test

In Section 4 of Unit 9 you learned how to calculate by hand the confidence intervals for population means that are based on the one-sample z-test. When the sample mean and sample standard deviation have already been calculated, these calculations are not long or tedious – so it does not save much time to calculate the confidence interval using Minitab instead. However, it is something that you have produced using Minitab already.

Computer activity 75 *Revisiting the testing of the average reading ability of seven-year-olds*

In Computer activity 64 (Subsection 7.3) you tested a hypothesis based on data from the BCS using a one-sample z-test. This hypothesis related to μ, the population mean of the British Ability Scales reading scores of all British 7-year-old children in 2004–2005. Summary statistics for the data were given in Table 2 (Subsection 7.3) and are repeated below for convenience.

Table 5 Summary statistics for the data on reading scores of 7-year-old children

Sample size	Sample mean	Sample standard deviation
396	111.28	26.668

(These data are copyright and owned by the Economic and Social Data Service.)

You may recall that in Computer activity 64 it was noted that part of the output corresponded to a confidence interval. This confidence interval is the 95% confidence interval for the population mean based on the z-test. So, in doing that activity, you have already learned how to produce a 95% confidence interval in Minitab.

It is possible to get Minitab to produce the output for the 95% confidence interval without the output for the one-sample z-test too. Produce this now for μ by doing the following.

- Obtain the **One-Sample Z for the Mean** dialogue box (**Stat** > **Basic Statistics** > **1-Sample Z**).
- Select `Summarized data` from the top drop-down list. Enter `396` in the **Sample size** field and `111.28` in the **Mean** field.
- Enter `26.668` in the **Known standard deviation** field.
- Make sure that **Perform hypothesis test** is *not selected*. The completed dialogue box is as in Figure 37.

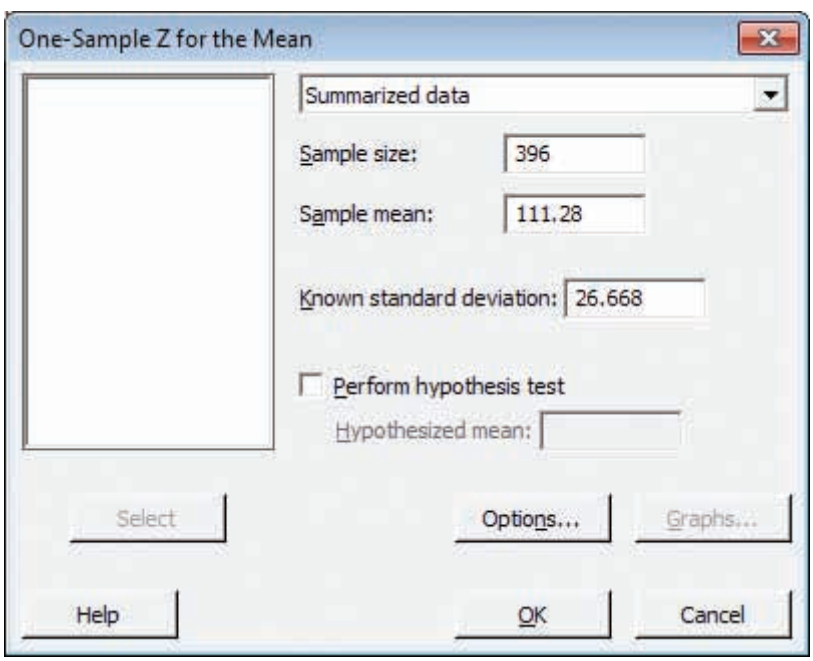


Figure 37 The **One-Sample Z for the Mean** dialogue box ready to calculate a confidence interval

- Click on **OK**.

The output produced by Minitab is as follows.

```
The assumed standard deviation = 26.668

   N     Mean   SE Mean         95% CI
396   111.28      1.34   (108.65, 113.91)
```

This output was also produced in Computer activity 64. However, in addition, the results of a hypothesis test were also given.

The 95% confidence interval is the entry beneath `95% CI`, which is (108.65, 113.91). So on the basis of the sample that we have, (108.65, 113.91) is the 95% confidence interval for μ, the population mean of the British Ability Scales reading scores of all British 7-year-old children in 2004–2005.

It is also possible to use Minitab to obtain the 99% confidence interval for μ based on the one-sample z-test, by doing the following.

- Obtain the **One-Sample Z for the Mean** dialogue box again (**Stat** > **Basic Statistics** > **1-Sample Z**).
- Make sure the dialogue box is set up in the same way as it was for the 95% confidence interval above, but do not click on **OK** yet.
- Click on **Options...** to obtain the **One-Sample Z: Options** dialogue box. Change the **Confidence level** field to 99. (In fact, any confidence interval for μ between 0.0001% and 99.9999% can be obtained by entering the corresponding percentage in the **Confidence level** field.)

  The completed dialogue box is as in Figure 38.

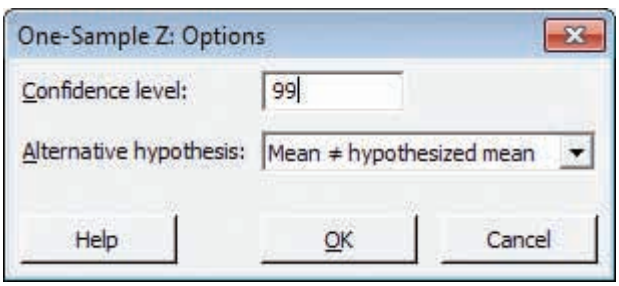


Figure 38 The **One-Sample Z: Options** dialogue box

- Click on **OK**, and then click on **OK** in the **One-Sample Z for the Mean** dialogue box.

Notice that the output now produced by Minitab gives the 99% confidence interval for μ: (107.83, 114.73). This interval is wider than the 95% confidence interval you produced before. However, both intervals are centred around 111.28 – the sample mean.

Practise using Minitab to obtain confidence intervals by doing the following activity.

Computer activity 76 *Confidence intervals for jam*

In Activity 18 of Unit 9 (Subsection 4.2) you calculated a 95% confidence interval for the mean weight of a particular manufacturer's plum jam. For the sample of 37 jars on which this calculation was based, the file **jam.mtw** gives the weights (in grams) of the plum jam.

(a) Using Minitab, obtain the mean and standard deviation of this sample.

(b) Also using Minitab, obtain a 95% confidence interval for the mean weight of such jars of plum jam.

(c) Again using Minitab, obtain a 99% confidence interval for the mean weight.

(d) On the label of the jam, it claims that the jar contains, on average, 454 g (1 lb). Is this claim reasonable on the basis of the confidence intervals you obtained in parts (b) and (c)?

9.3 Exploring intervals from fitted lines

This subsection focuses on the interpretation of confidence intervals for the mean response. In the following activities, you will use a couple of interactive computer resources to investigate how such intervals vary when different samples are taken from the population.

Computer activity 77 *Effect of sampling on the least squares fit line*

Open the interactive computer resource 'Sampling lines' on the M140 website.

Initially, this resource displays a line summarising the relationship between two variables, x and y, in a population.

(a) Generate a sample and compare the least squares regression line with the population line.

(b) Generate another sample. Is the least squares regression line based on this new sample any closer to the population line?

(c) Now generate 100 samples and compare the resulting lines. What do you notice?

Computer activity 78 *Repeated samples*

In Computer activity 77, you considered the sampling variability of the least squares regression lines. In this activity, you are going to concentrate on one particular value of x: $x = 5$.

Open the interactive computer activity 'Confidence intervals' on the M140 website.

(a) Generate one sample and note down the confidence interval. What value is the interval centred around, and what is the width of the interval? (That is, what are $(y_{\min} + y_{\max})/2$ and $y_{\max} - y_{\min}$?)

(b) For these data, the population slope happens to be 0.5 and the population intercept is 2.5. Calculate the value of y_{true} when $x = 5$. Did the interval you obtained in part (a) contain this value?

(c) Generate another sample, and note down this confidence interval too. Does this interval match the one you obtained in part (a)? Does this interval contain y_{true}?

(d) Click on 'Reset' and then generate 100 samples. How many of your intervals contain the value of y_{true}?

Computer activity 79 *Confidence intervals and scatter*

In Computer activity 78, for points scattered around a particular line you saw that roughly 95% of the confidence intervals for a mean response contained the correct value (that is, the value of y that is given by the population line). In that activity, the strength of the relationship between the x and y values was roughly the same in each sample. In this activity, you will explore the impact that the strength of the relationship has on the confidence intervals.

Open the interactive computer resource 'Confidence intervals', if it is not already open, and click on 'Reset'.

(a) Generate samples with different strengths of relationship. What do you notice about the scatter of the points about the line as the strength of the relationship alters?

(b) Generate a sample for which the relationship is weak, and note down the 95% confidence interval. Do the same for moderate and strong relationships, and compare the widths of the three intervals. What do you notice?

(c) Now generate 100 samples where, for each of them, the relationship is weak. Note how many times the 95% confidence interval contains y_{true}. Then do the same for moderate and strong relationships.

Computer activity 80 *Confidence intervals and population slope*

In Computer activity 79, you saw that, although the width of a confidence interval depends on the strength of the relationship, the chance that each one contains the population value of y remains the same.

In this activity, you will explore the impact of the slope of the population line. So first make sure that the interactive computer resource 'Confidence intervals' is open, and click on 'Reset'.

(a) Generate three samples when the population slope is 0.3 (with the strength of the relationship set to 'moderate'). Calculate the width of the confidence interval when $x = 5$ in each case.

(b) Again with the population slope set to the value 0.3 (and with the strength of the relationship set to 'moderate'), generate 100 samples. How many confidence intervals contain the correct value when $x = 5$ (that is, the value of y given by the population line when $x = 5$)?

(c) Repeat parts (a) and (b), with the population slope set to the value 10.

(d) Again repeat parts (a) and (b), but this time using a population slope of -5.

(e) Compare your results for parts (a) to (d). Does the population slope make a difference?

Computer activity 81 *Confidence intervals and sample size*

In Computer activities 78 to 80, each line was based on a sample of 10 points. In this activity you will explore the effect of changing the sample size.

Make sure that the interactive computer resource 'Confidence intervals' is open, and click on 'Reset'.

(a) Generate 100 samples of 10 points. How many contain the correct (i.e. population) value for y when $x = 5$? For your last generated sample, how wide is the confidence interval for $x = 5$?

(b) Generate 100 samples of 40 points. How many contain the correct value when $x = 5$? For your last generated sample, how wide is the confidence interval for $x = 5$?

(c) Generate 100 samples of 160 points. How many contain the correct value when $x = 5$? For your last generated sample, how wide is the confidence interval?

9.4 Obtaining confidence intervals and prediction intervals for fitted lines

In Computer activities 82 and 83, you will learn how to use Minitab to obtain confidence intervals for the mean response and prediction intervals.

Computer activity 82 *Obtaining a confidence interval for the mean response*

Activity 25 of Unit 9 (Subsection 5.1) gave the 95% confidence interval as 48.6% to 55.2% for the mean pass rate in Mathematics for schools with a pass rate in English of 50%. In Example 16 of Unit 9 (Subsection 5.2), the corresponding 95% prediction interval was given as 33.6% to 70.3%.

The data on which this confidence interval was based are given in the file **ks4subjects.mtw**. Open this worksheet in Minitab now.

Confidence intervals for the mean response and predictions intervals are obtained only after the regression line has been fitted using Minitab. So first do the following.

- Click on **Stat**, then **Regression**, then **Regression** and then **Fit Regression Line...**. This opens the **Regression** dialogue box.
- A confidence interval and prediction interval for the pass rate in Mathematics is required. This means that the response variable is `maths` and the explanatory variable is `english`.

- So enter `maths` in the **Responses** field and enter `english` in the **Continuous predictors** field.
- Click on **OK**.

This ensures that the regression line has been fitted in Minitab. Now, to get Minitab to calculate confidence intervals and prediction intervals based on this regression line, do the following.

- Click on **Stat**, then **Regression**, then **Regression** and then **Predict...**. This opens the **Predict** dialogue box.
- The confidence interval for the mean response and the prediction interval for the Mathematics pass rate when the English pass rate is 50% are required. So enter `50` in the **english** field. The completed dialogue box is as in Figure 39.

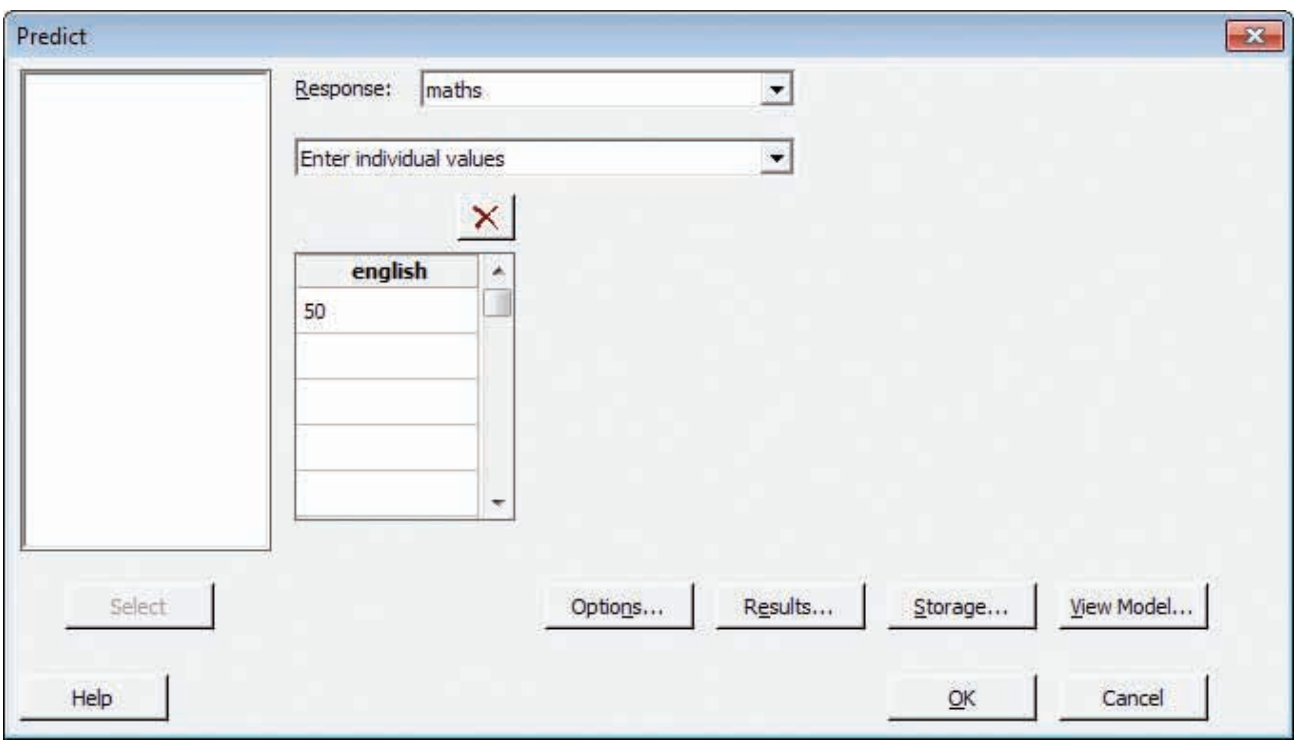


Figure 39 The **Predict** dialogue box

- Click on **OK**.

Look in the Session window. The ends of the 95% confidence interval are given in brackets under `95% CI`: (48.6270, 55.1805), that is, to one decimal place, 48.6% to 55.2%.

The ends of the 95% prediction interval are given under `95% PI`: (33.5539, 70.2535). That is, 33.6% to 70.3% to one decimal place.

Computer activity 83 *Predicting performance at Key Stage 4 in Welsh constituencies*

In Activity 14 of Unit 9 (Subsection 3.3), data on the percentage of children achieving particular benchmarks at Key Stage 2 in 2006 and Key Stage 4 in 2011, for eight constituencies in South Wales East, were plotted. These data are given in the file **South Wales East KS2 and KS4 results.mtw**. Open this file in Minitab.

Note that the results at Key Stage 2 are given in the variable `KS2` and the results at Key Stage 4 are given in the variable `KS4`.

(a) In the following parts of this activity you are going to be asked to obtain confidence intervals for the mean response and prediction intervals for the percentage of students reaching the benchmark at Key Stage 4 in 2011. So, using Minitab, first find the equation of the corresponding least squares regression line.

(b) In Blaenau Gwent, 68.0% of Key Stage 2 students reached the specified benchmark in 2006. Obtain and interpret the 95% confidence interval for the mean percentage of students reaching the benchmark at Key Stage 4 in 2011 for constituencies like Blaenau Gwent (that is, constituencies where 68.0% of Key Stage 2 students in 2006 reached the specified benchmark).

(c) In Caerphilly, 72.5% of Key Stage 2 students reached this benchmark in 2006. Obtain and interpret the 95% prediction interval for the percentage of students reaching the benchmark at Key Stage 4 in 2011 for a constituency like Caerphilly (that is, a constituency where 72.5% of Key Stage 2 students in 2006 reached the specified benchmark).

(d) Confidence intervals for the mean response, and prediction intervals for `KS4` based on all the values given in the variable `KS2`, can be obtained by selecting `Enter columns of values` in the second drop-down field in the **Predict** dialogue box and entering `KS2` in the **KS2** field. Do this now to give the confidence intervals and the prediction intervals in the Data window. The two ends of the confidence intervals are given in columns `CLIM1` and `CLIM2`. The two ends of the prediction intervals are given in the columns `PLIM1` and `PLIM2`. Compare the 95% prediction intervals for `KS4` with the actual values for each of the constituencies. Do the prediction intervals look reasonable?

Summary of Chapter 9

In this chapter, you have seen that confidence intervals based around fitted lines can be interpreted in terms of repeated samples. You have also seen that confidence intervals are narrower when the data are less scattered and when the sample is bigger.

You have learned how to use Minitab to calculate correlation coefficients. Finally, you have seen how Minitab can be used to obtain confidence and prediction intervals.

10 Experiments

In this chapter, which is associated with Unit 10, you will learn how to do the following hypothesis tests using Minitab: the one-sample t-test, the two-sample t-test and the paired t-test. As you will discover when doing these tests, Minitab will also provide corresponding confidence intervals.

10.1 One-sample t-test

In Subsection 7.3 of this Computer Book, you learned how to do the one-sample z-test using Minitab. The procedure for the one-sample t-test works in a very similar way – the major difference is which dialogue box needs to be selected.

Computer activity 84 *Testing the yield from a variety of tomatoes*

In Subsection 4.1 of Unit 10, you learned how to use the one-sample t-test to test the hypotheses

H_0: $\mu = 4$
H_1: $\mu \neq 4$,

where μ is the yield (in kg per plant) of one variety of outdoor bush tomatoes grown using a new fertiliser.

Yields from five tomatoes are given in the file **tomatoes.mtw**. Test the null hypothesis using a one-sample t-test by doing the following.

- Open the worksheet **tomatoes.mtw**.
- Click on **Stat**, choose **Basic Statistics**, and choose **1-Sample t...**. This opens the **One-Sample t for the Mean** dialogue box.

- It is possible to enter summary data (the sample size, mean and standard deviation) into the dialogue box in the same way that you did for the one-sample z-test. However, this means first working them out from the data in the worksheet. Instead, it is possible to work directly from the raw data. Do this now by making sure `One or more samples, each in a column` is selected in the drop-down list and entering `yield` in the corresponding field.
- Select **Perform hypothesis test**. Then enter `4` in the **Hypothesized mean** field, as this is the value of the mean assumed by the null hypothesis. The completed dialogue box should be as in Figure 40.

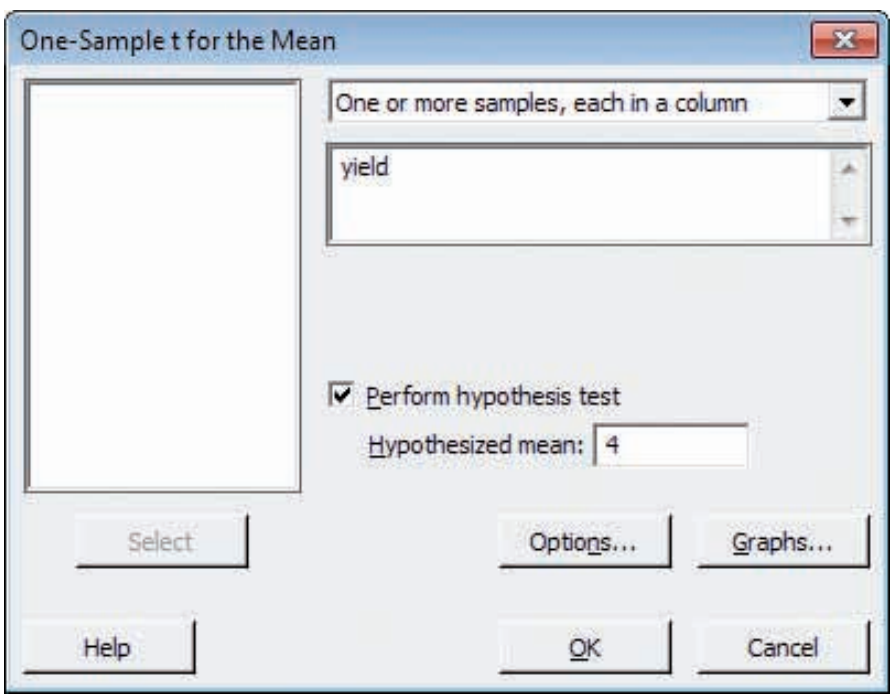


Figure 40 The **One-Sample t for the Mean** dialogue box

- Click on **OK**.

The output from the test produced by Minitab is as follows.

```
Test of μ = 4 vs ≠ 4

Variable  N   Mean  StDev  SE Mean      95% CI           T      P
yield     5  3.280  0.497    0.222  (2.663, 3.897)  -3.24  0.032
```

Most of the output is the same as Minitab would produce for the one-sample z-test. The null and alternative hypotheses are re-stated, and the sample size, mean and standard deviation are given – reassuringly these are the same values as were given in the solution to Activity 16 in Unit 10. The estimated standard error (ESE) – `SE Mean` – is given as `0.222`, which matches that given in Example 11 of Unit 10.

Also in the output is the value `-3.24` for the test statistic, `T`, and the value `0.032` for the corresponding p-value, `P`. Using Table 1 in Subsection 6.2, this p-value represents moderate evidence against the null hypothesis, that is, moderate evidence that the yield of this variety of tomato is not 4 kg per plant when the new fertiliser is used. (In fact, the yield appears to be less than 4 kg.)

Notice that as part of the output for the t-test, Minitab provides a 95% confidence interval for the mean yield (based unsurprisingly on the t-test, not the z-test). This interval, `(2.663, 3.897)`, is entirely below the hypothesised value of 4. So, based on the data, a plausible range for the mean yield is (2.66 kg, 3.90 kg) per plant.

In Computer activity 84, you saw that Minitab automatically calculates a 95% confidence interval for the population mean based on the t-test. In Minitab it is possible to obtain the confidence interval without the output for the t-test. This is done by making sure that the **Perform hypothesis test** option is *not* selected in the **One-Sample t for the Mean** dialogue box. Note that when this is done, it does not matter if a number has been entered in the **Hypothesized mean** field.

The next computer activity gives you some practice in doing the one-sample t-test and obtaining the corresponding 95% confidence interval in Minitab.

Computer activity 85 *Measuring the speed of light*

The file **lightspeed.mtw** contains 23 measurements of the speed of light in air, made by Albert Michelson in 1882. (Data source: Stigler, S.M. (1977) 'Do robust estimators work with real data?', *Annals of Statistics*, vol. 5, pp. 1055–98.) Each of the measurements is in km/s over 299 000. It is reasonable for you to assume that the measurements come from a normal population distribution, because of the way they were taken.

(a) Using Minitab, obtain the 95% confidence interval for the speed of light in air.

(b) According to one source, the speed of light in air is 299 705 km/s. This corresponds to the hypothesis

$$H_0\colon\ \mu = 705$$
$$H_1\colon\ \mu \neq 705,$$

where μ is the speed of light in air in km/s over 299 000. Test this claim using a one-sample t-test. What value of the test statistic do you obtain, and what is the p-value? Hence what conclusion do you draw?

(c) Does the confidence interval you obtained in part (a) match the conclusion you drew in part (b)? Why or why not?

10.2 Two-sample t-test

In Subsection 10.1, you learned how to do the one-sample t-test in Minitab. Using Minitab it is also possible to do the two-sample t-test, as you will discover in this subsection.

Computer activity 86 *Ball manoeuvres*

In Activities 10 and 12 of Unit 10 (Subsection 3.3) you tested whether the method of preparation made a difference to the length of time a child took to manoeuvre a ball round an obstacle course and into a hole. Those activities involved quite a bit of calculation by hand. Use Minitab to complete those same calculations by doing the following.

- Open the worksheet **obstaclecourse.mtw**. Notice in the worksheet that the data for the two groups are given in separate columns. This is not the only way data for the two-sample t-test can be structured in Minitab, but it is the way we will use for now.
- Obtain the **Two-Sample t for the Mean** dialogue box by clicking on **Stat**, then **Basic Statistics**, and then **2-Sample t...**.
- In the dialogue box select `Each sample is in its own column` from the top drop-down list. Then in the **Sample 1** field enter `A` and in the **Sample 2** field enter `B`. The completed dialogue box is shown in Figure 41.
- By default Minitab performs a version of the two-sample t-test that does not make the assumption of a common population variance. To ensure that Minitab does the same version of the two-sample t-test that is described in Unit 10, click on the **Options...** button in the **Two-Sample t for the Mean** dialogue box to bring up the **Two-Sample t: Options** dialogue box. In this dialogue box tick the **Assume equal variances** option.

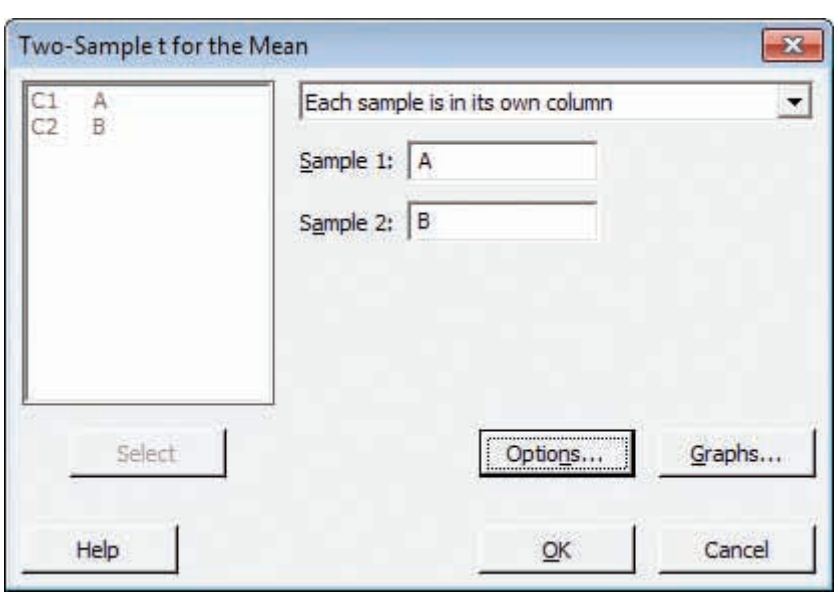


Figure 41 The **Two-Sample t for the Mean** dialogue box

- Click on **OK** to return to the **Two-Sample t for the Mean** dialogue box, and then on **OK** again.

The output produced by Minitab is as follows.

```
Two-sample T for A vs B

   N  Mean  StDev  SE Mean
A  5  5.00   2.55      1.1
B  5  7.00   3.08      1.4


Difference = μ (A) - μ (B)
Estimate for difference:  -2.00
95% CI for difference:  (-6.13, 2.13)
T-Test of difference = 0 (vs ≠): T-Value = -1.12  P-Value = 0.296  DF = 8
Both use Pooled StDev = 2.8284
```

As with the one-sample t-test, Minitab starts by giving details about the test it has just done and gives some summary statistics. The output also automatically gives the 95% confidence interval for the difference between the two means. (In this case, the output makes it clear that the difference is taken to be $\mu_A - \mu_B$.) The last line gives the value of the pooled standard deviation – reassuringly the same as the value calculated in Activity 10 of Unit 10 (Subsection 3.3).

The value of the test statistic, t, and the associated p-value are given in the penultimate line of the output. The value $t = -1.12$ is the same as that calculated in Activity 12 of Unit 10, and a p-value of 0.296 matches the conclusion drawn in that activity – namely, there is little evidence that how the children were prepared for the obstacle course affected the length of time they took to negotiate it.

Computer activity 87 *Effect of selection*

The file **fruitfly.mtw** contains data on the egg-laying capability of female fruit flies of the species *Drosophila melangoster*. (Data source: Sokal, R.R. and Rohlf, F.J. (1995) *Biometry: The Principles and Practice of Statistics in Biological Research*, 3rd edn, W.H. Freeman and Company.) Each data value is the number of eggs laid per female per day in the first 14 days of life. The females in the `selected` group have been selectively bred to change their resistance to DDT (an insecticide); the females in the `control` group have not been selectively bred.

(a) Produce a single diagram that contains boxplots of the number of eggs laid per female per day for each of the two groups (**Graph** > **Boxplot**).

From the boxplots, is it reasonable to assume that in both groups the population distribution of the number of eggs laid per female per day is normal?

(b) Using the two-sample t-test, test the hypothesis that the selective breeding did not alter the egg-laying capability of this species of fruit fly.

(c) Is it reasonable to have used a pooled standard deviation in the test you did in part (b)?

(d) Give a 95% confidence interval for $\mu_s - \mu_c$, where μ_s is the egg-laying capability of selectively bred fruit flies and μ_c is the egg-laying capability of fruit flies that have not been selectively bred.

10.3 Matched-pairs t-test

Subsection 4.2 of Unit 10 focused on the matched-pairs t-test. There you learned that this t-test can be thought of as a particular form of the one-sample t-test – that is, a one-sample t-test where the data consist of differences (d), and the hypotheses are

$$H_0\colon\ \mu_d = 0$$
$$H_1\colon\ \mu_d \neq 0.$$

In this subsection you will learn how to do this test using Minitab, without having to calculate the differences explicitly.

Computer activity 88 *Testing the effect on sleep of two forms of a drug*

In Subsection 4.2 of Unit 10, you considered whether two different forms, L and R, of a sleep-inducing drug, hyoscyamine hydrobromide, differ in their capacity to induce sleep. The data are given in **bromide.mtw**. Use Minitab to test the hypotheses

$$H_0\colon\ \mu_{L-R} = 0$$
$$H_1\colon\ \mu_{L-R} \neq 0,$$

where μ_{L-R} is the difference between sleep gain (in hours) when the L form of the drug was taken compared with when the R form of the drug was taken, by doing the following.

- Open the worksheet **bromide.mtw**. Notice how the data are structured in the worksheet. The data relating to each form of the drug are given in a separate column, and each row gives the data for a particular patient.
- Obtain the **Paired t for the Mean** dialogue box by clicking on **Stat**, then **Basic Statistics**, and then **Paired t...**.
- In the **Paired t for the Mean** dialogue box, make sure `Each sample is in a column` is selected from the top drop-down list. Then in the **Sample 1** field enter `formL` and in the **Sample 2** field enter `formR`. The completed dialogue box is shown in Figure 42.

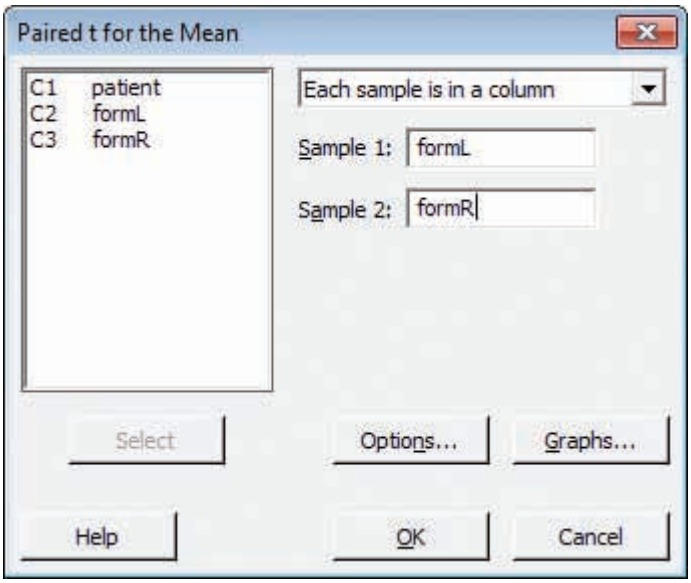


Figure 42 The **Paired t for the Mean** dialogue box

- Click on **OK**.

The output produced by Minitab is as follows.

```
Paired T for formL - formR

               N   Mean  StDev  SE Mean
formL         10  2.330  2.002    0.633
formR         10  0.750  1.789    0.566
Difference    10  1.580  1.230    0.389


95% CI for mean difference: (0.700, 2.460)
T-Test of mean difference = 0 (vs ≠ 0): T-Value = 4.06  P-Value = 0.003
```

Notice that this output is structured in a similar way to the output for the one-sample t-test and the two-sample t-test. Some summary statistics are given, along with a 95% confidence interval and details about the test, particularly the value of the test statistic and the corresponding p-value.

So for these data and hypotheses, the test statistic is 4.06 and the p-value is 0.003. This means that there is strong evidence that the sleep gain is not the same for the two forms of hyoscyamine hydrobromide. In fact, it appears that the L form of the drug induces more sleep. An indication of just how much more sleep the L form of the drug induces is provided by the 95% confidence interval. The interval $(0.700, 2.460)$ means that it is plausible that the L form of the drug induces anywhere between 0.700 and 2.460 more hours of sleep than the R form of the drug.

Computer activity 89 *Comparing measuring devices*

In a study to compare measuring devices, readings from 10 samples were taken using two measuring devices, device A and device B. Of interest is whether it matters which machine is used to take the readings – that is, whether on average they both give the same value. This corresponds to the hypotheses

$$H_0\colon\ \mu_{A-B} = 0$$
$$H_1\colon\ \mu_{A-B} \neq 0,$$

where μ_{A-B} is the population mean difference in readings taken on device A compared with device B. The data are given in the file **devices.mtw**. (Data source: Hahn, G.J. and Nelson, W. (1970) 'A problem in the statistical comparison of measuring devices', *Technometrics*, vol. 12, pp. 95–102.)

(a) Briefly explain why the matched-pairs t-test is suitable for these data. (You may assume that differences in readings are normally distributed.)

(b) Using a matched-pairs t-test, test the null hypothesis. Clearly state your conclusions.

(c) State a 95% confidence interval for the difference in readings given by device A compared with device B.

10.4 One-sided t-tests and z-tests

So far in this chapter, you have just been doing two-sided t-tests using Minitab. That is, the alternative hypothesis H_1 has been that a population mean, or difference between population means, is *not equal* to a stated value.

However, as you learned in Section 6 of Unit 10, sometimes a one-sided alternative hypothesis is more appropriate – that is, an alternative hypothesis of the form

$$H_1\colon\ \mu < A \quad \text{or} \quad H_1\colon\ \mu > A,$$

for a one-sample t-test or one-sample z-test, and of the form

$$H_1\colon\ \mu_A < \mu_B \quad \text{or} \quad H_1\colon\ \mu_A > \mu_B,$$

for a two-sample t-test.

It is possible to do one-sided tests in Minitab as well as two-sided ones. In fact, the means by which a one-sided test is obtained instead of the two-sided version is the same for the one-sample t-test, two-sample t-test, matched-pairs t-test and one-sample z-test. To see how it is done for one of these tests we revisit an example introduced in Unit 10.

Computer activity 90 *Benefit of exercise*

In Activity 26 of Unit 10 (Section 6) some data were given on resting heart rate before and after a one-year exercise programme. As explained in the solution to Activity 26, the following hypotheses are appropriate:

H_0: $\mu_d = 0$

H_1: $\mu_d < 0$,

where μ_d denotes the population mean change in resting heart rate over the course of the exercise programme. So a one-sided test is required. The data, which are given in the file **exercise.mtw**, consist of pairs of measurements for seven people. So this means that, in particular, a one-sided matched-pairs t-test is required.

To perform the test in Minitab, start in precisely the same way as you did for a two-sided matched-pairs t-test in Subsection 10.3, as follows.

- Open the worksheet **exercise.mtw**.
- Click on **Stat**, choose **Basic Statistics** and then **Paired t...**. The **Paired t for the Mean** dialogue box appears.
- Make sure that `Each sample is in a column` is selected from the drop-down list. In the **Sample 1** field enter `after` and in the **Sample 2** field enter `before`. (In this case, `after` is chosen to be the first sample so that the differences taken correspond to `after` − `before`.)

Now there is one extra step to make sure that Minitab does a one-sided rather than two-sided test.

- Click on **Options...** in the **Paired t for the Mean** dialogue box. The **Paired t: Options** dialogue box appears.
- The alternative hypothesis corresponds to H_1: $\mu_d < 0$, so in the **Paired t: Options** dialogue box select `Difference < hypothesized difference` in the **Alternative hypothesis** field. The completed dialogue box is shown in Figure 43.

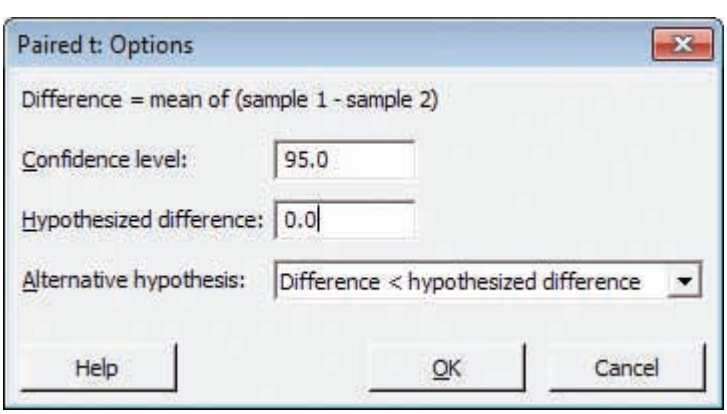


Figure 43 The **Paired t: Options** dialogue box

- Click on **OK** to return to the **Paired t for the Mean** dialogue box, and click on **OK** again.

The resulting Minitab output is shown below.

```
Paired T for after - before

            N    Mean  StDev  SE Mean
after       7   69.57   2.64     1.00
before      7   72.00   2.83     1.07
Difference  7  -2.429  1.813    0.685


95% upper bound for mean difference: -1.097
T-Test of mean difference = 0 (vs < 0): T-Value = -3.54  P-Value = 0.006
```

As before with the matched-pairs t-test, the standard deviation (`StDev`), sample size (`N`), sample mean (`Mean`) and ESE (`SE Mean`) are quoted for both variables and for the difference. However, notice this time that the statement of the alternative hypothesis is `< 0` rather than `not = 0`, confirming that Minitab has indeed done a one-sided test. A form of a confidence interval appropriate to the one-sided test (`95% upper bound`) is then given. In the solution to Activity 26 in Unit 10, the test statistic t was given as -3.545, which matches (apart from rounding) the value of t (`T-Value`) given by Minitab. Finally, the p-value given by Minitab (`P-Value`) is 0.006. So, from Table 1 (Subsection 6.2), this means there is strong evidence against the null hypothesis. That is, there is strong evidence that the population mean resting heart rate is lower after the exercise programme compared with the resting heart rate before the exercise programme.

The dialogue boxes you have used to perform the one-sample z-test, the one-sample t-test and the two-sample t-test also have an **Options...** button. Clicking on this button generates a dialogue box similar to the **Paired t: Options** dialogue box, where you can change the entry in the **Alternative hypothesis** field to switch between two-sided and one-sided versions of these other tests. (In Subsection 9.2, you have already used this dialogue box to specify which confidence interval from a z-test you wish Minitab to calculate, and in Subsection 10.2 you used this dialogue box to ensure that Minitab carried out the right version of the two-sample t-test.)

Computer activity 91 provides practice in carrying out one of the other one-sided tests using Minitab.

Computer activity 91 *Do people in pain sleep for fewer hours a night?*

An American study of the sleeping problems associated with chronic widespread pain (fibromyalgia) considered many aspects, just one of which was the simple question: *Do people with fibromyalgia sleep for fewer hours a night than people without?* According to the study of a random sample of 744 American fibromyalgia patients, their mean sleeping time was 5.6 hours per night, with standard deviation 1.6 hours. Interest was in whether the population mean sleeping time of American fibromyalgia sufferers was less than the 'normative value' of 6.8 hours per night for the entire population of the USA.

(Source: Cappelleri, J.C. et al. (2009) 'Measurement properties of the Medical Outcomes Study Sleep Scale in patients with fibromyalgia', *Sleep Medicine*, vol. 10, pp. 766–70.)

(a) Write down the null and alternative hypotheses.

(b) Explain why it is appropriate to use a one-sample z-test to test the hypotheses you wrote down in part (a).

(c) Use a one-sided one-sample z-test to test the hypotheses you wrote down in part (a). Comment on your result.

Summary of Chapter 10

In this chapter, you have learned how to do various t-tests in Minitab: the two-sample t-test, the one-sample t-test and the matched-pairs t-test. The output from all these tests includes corresponding 95% confidence intervals, thus providing a way to obtain these intervals using Minitab. You have also learned how to switch between doing two-sided and one-sided versions of these tests and also between two-sided and one-sided versions of the one-sample z-test.

11 Clinical trials

This chapter covers two practical elements associated with Unit 11. First, Subsection 11.1 demonstrates most of the randomisation methods for clinical trials that were discussed in Subsection 3.4 of Unit 11. Second, Subsection 11.2 uses the χ^2 test and a t-test to analyse some data from clinical trials. This revises what you learned in Units 8 and 10.

11.1 Randomisation in practice

In this subsection you will explore different schemes for randomising patients in a clinical trial. In Computer activities 93 to 95 you will use Minitab to assign patients to treatments systematically, randomly, and then randomly but keeping the totals fixed.

By its very nature, randomisation leads to unpredictable results – at least in terms of which subjects get assigned to each group. It is, however, possible to obtain exactly reproducible results by exploiting the means by which Minitab generates random numbers.

Random numbers produced by Minitab are in fact 'pseudo-random numbers' – that is, the numbers follow a predetermined sequence, but in such a way that they appear random. This means that by starting the sequence at the same place every time – 'setting the seed' – the same sequence of pseudo-random numbers will always be obtained. So first, in the next activity, you will learn how to set the seed for the random number generator in Minitab.

Computer activity 92 *Setting the seed*

(a) In a new worksheet in Minitab (**File** > **New**) use the same procedure as you did in Computer activity 40 (Subsection 4.2) to obtain a column of 30 numbers between 1 and 100, stored in column `C1` (**Calc** > **Random Data** > **Integer**).

(b) Now set the seed of the random number generator to take the value 10 by doing the following. (The value 10 just happened to be the value chosen by the module team. It has no significance other than it's a value chosen by somebody.)

- Click on **Calc** and then **Set base...** to open the **Set Base** dialogue box.
- Enter `10` in the **Set base of random data generator to** field. The completed dialogue box is shown in Figure 44.

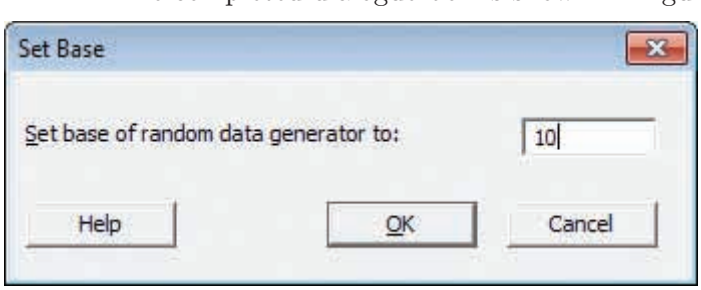


Figure 44 The **Set Base** dialogue box

- Click on **OK**.

Generate another column of 30 random numbers between 1 and 100, this time storing them in column `C2`.

(c) Repeat part (b), this time storing the column of random numbers in column `C3`.

(d) Generate a fourth column of 30 random numbers between 1 and 100, storing them in column `C4`. (This time do *not* reset the seed first.)

Suppose a clinical trial consisting of 100 participants spread over two centres is being set up, and these participants need to be randomised into the experimental and control groups. Open the worksheet **randomisation.mtw**, which contains two variables `patientID` and `centre`. The variable `patientID` is a patient identifier, and patients are simply numbered 1 to 100. The variable `centre` identifies the centre that the patient will be attending for the trial; there are two centres, numbered 1 and 2. Assignments to groups are to be numbered so that `0` indicates 'assign to the control group', and `1` indicates 'assign to the experimental group'.

In each of the following activities you will assign participants to treatment groups. Note that the solution for each activity assumes that the seed for the random number generator is set to take the value 42.

Computer activity 93 *Assigning patients systematically*

In this activity, you will assign trial participants to groups in a systematic way: odd patient IDs are assigned to the control group and even patient IDs are assigned to the experimental group.

- Click on **Calc**, then **Make Patterned Data** and then **Simple Set of Numbers...**. The **Simple Set of Numbers** dialogue box will now appear.
- Enter `treatment` in the **Store patterned data in** field.
- Enter `0` in the **From first value** field.
- Enter `1` in the **To last value** field.
- Amend **Number of times to list the sequence** to `50`.
- Click on **OK**.

Create a contingency table of the number of participants in each of the treatment groups in each centre by doing the following.

- Click on **Stat**, then **Tables** and then **Cross Tabulation and Chi-Square...** to obtain the **Cross Tabulation and Chi-Square** dialogue box.
- Make sure that `Raw data (categorical variables)` is selected in the top drop-down list. In the **Rows** field, enter `centre`, and in the **Columns** field, enter `treatment`.
- Make sure that the **Counts** option is selected. The completed dialogue box is shown in Figure 45.

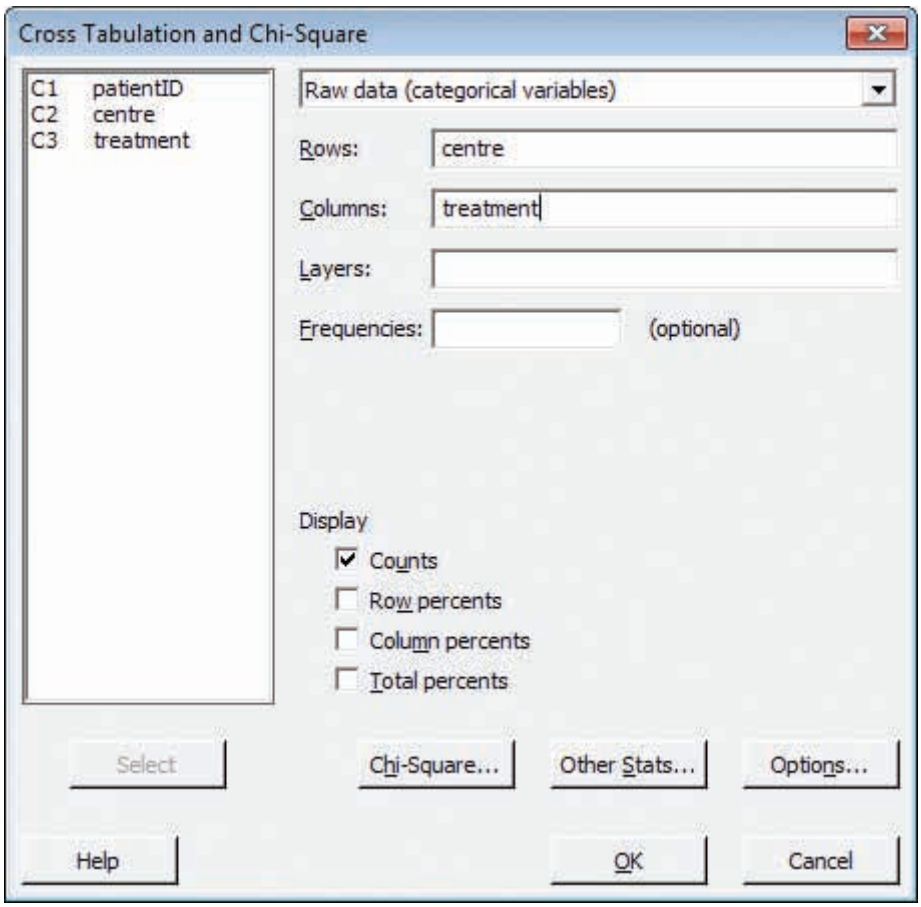


Figure 45 The **Cross Tabulation and Chi-Square** dialogue box

- Click on **OK**.

Notice that when you do this, a contingency table appears in the Session window giving the number of participants assigned to each treatment group in each centre. The rows in this table correspond to centres and the columns correspond to treatments.

(a) Overall, how many participants were assigned to each treatment group?

(b) Is there balance in the number of participants assigned to each centre?

(c) If the seed for the random number generator is not set, will the assignment change?

Computer activity 94 *Assigning patients randomly*

In this activity, you will assign trial participants to treatment groups at random – with probability 0.5 of being assigned to either group, as if tossing a coin.

- Make sure that the worksheet **randomisation.mtw** is the active worksheet.
- If you want to get exactly the same answers as given in our solution, set the seed for the random number generator to the value 42 (**Calc** > **Set base**).
- Click on **Calc**, then **Random Data** and then **Binomial...**. The **Binomial Distribution** dialogue box will now appear.
- Enter `100` in the **Number of rows of data to generate** field.
- Enter `treatment` in the **Store in column(s)** field.
- For each participant, the outcome is like the outcome from a single coin toss. So enter `1` in the **Number of Trials** field.
- Enter `0.5` in the **Event probability** field. The completed dialogue box is shown in Figure 46.

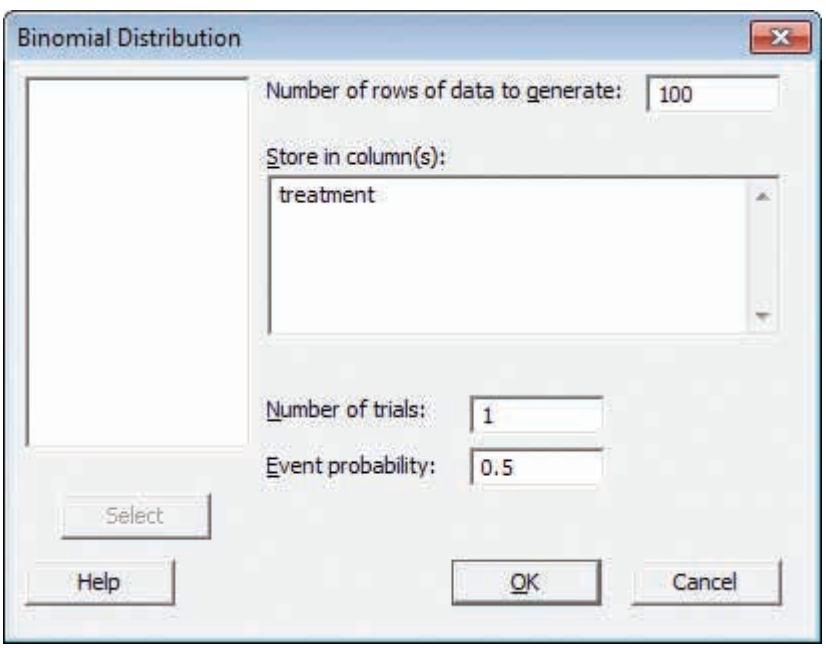


Figure 46 The **Binomial Distribution** dialogue box

- Click on **OK**.

Your randomisation should appear in the column labelled `treatment`. Create a contingency table of the numbers of participants assigned to each group at each centre (**Stat** > **Tables** > **Cross Tabulation and Chi-Square**).

What can you say about the balance of the groups?

Computer activity 95 *Assigning patients randomly but keeping the totals fixed*

You will now assign trial participants by fixing the total in each group (50 in each), but ordering randomly. This is done by first assigning treatments to participants systematically, and then shuffling these assignments.

- Make sure that the worksheet **randomisation.mtw** is the active worksheet.
- If you want to get exactly the same answers as given in our solution, set the seed for the random number generator to the value 42 (**Calc** > **Set base**).
- Create a column `treatment` in which participants are assigned to treatment groups systematically (**Calc** > **Make Patterned Data** > **Simple Set of Numbers**). (This was done in Computer activity 93.)
- Click on **Calc**, then **Random Data** and then **Sample From Columns...**. The **Sample From Columns** dialogue box will now appear.
- In Subsection 4.2, you used this same dialogue box to select a sample from a population. If a sample is specified to be the same size as the population, this results in the 'sample' just being a randomly shuffled copy of the 'population'. So in the **Sample From Columns** dialogue box, enter `100` in the **Number of rows to sample** field, `treatment` in the **From columns** field and `treatment` in the **Store samples in** field.
- Click on **OK**.

Your randomisation will appear in the column labelled `treatment`. Create a contingency table of the numbers of participants assigned to each group at each centre (**Stat** > **Tables** > **Cross Tabulation and Chi-Square**).

What can you say about the balance of the groups?

In Computer activity 95, you randomly assigned participants to treatment groups in such a way that the total number of participants in each treatment group was fixed. However, as you saw, this did not guarantee that the allocation was balanced. To achieve this, a stratified randomisation is required. There are different ways of achieving this using Minitab; however, the most straightforward way is just to do the same as you did in Computer activity 95 – but deal with each centre separately. We will not ask you to do stratified randomisation in M140.

This subsection has demonstrated most of the randomisation schemes described in Subsection 3.4 of Unit 11, which would be suitable for a group-comparative design. Matched-pairs and crossover designs require randomisation to be done in sets of pairs, but the procedure is similar. Remember that forcing balance may only be required when numbers

within each stratum are small. Packages are available that allow stratified randomisation to be carried out much more easily than in Minitab.

11.2 Analysis of data from clinical trials

In this subsection you will use Minitab to analyse data from a couple of clinical trials.

Computer activity 96 *A new treatment for pneumonia?*

A phase 3 clinical trial was carried out to test the effectiveness and safety of a new antibiotic drug, ceftaroline, against an existing antibiotic drug, ceftriaxone, in curing patients with pneumonia. Patients hospitalised with pneumonia were randomised to the experimental group – 613 received ceftaroline – or the control group – 615 received ceftriaxone. Participants who could be evaluated for a cure had to have received their treatment for between two and seven days, and to have completed a 'test of cure' visit (8 to 15 days after hospitalisation) with results that could be evaluated: 459 patients were evaluated for ceftaroline and 449 for ceftriaxone. The main effectiveness outcome was cure: bacteria that caused pneumonia in the patient were eliminated at the time of the test of cure visit. Data on safety outcomes were also published, and the most common reported adverse event in this study was diarrhoea.

(Source: File, T.M. et al. (2010) 'Integrated analysis of FOCUS 1 and FOCUS 2: randomized, doubled-blinded, multicenter phase 3 trials of the efficacy and safety of ceftaroline fosamil versus ceftriaxone in patients with community-acquired pneumonia', *Clinical Infectious Diseases*, vol. 51, issue 12, pp. 1395–405.)

These data are given in the Minitab worksheet **ceftaroline.mtw**. Open this worksheet. Notice the worksheet contains columns that relate to two different contingency tables. One contingency table, consisting of the columns `treatment`, `cure` and `nocure`, relates to the data on the effectiveness of ceftaroline. The other contingency table, consisting of the columns `treatment`, `diarrhoea` and `nodiarrhoea`, relates to the data on the safety of ceftaroline.

(a) Why is the χ^2 test an appropriate hypothesis test to be using with either of these contingency tables?

(b) Use Minitab to a carry out a χ^2 test to test the following null hypothesis:

H_0: No difference in pneumonia cure rate in patients taking ceftaroline and ceftriaxone.

(c) Use Minitab to a carry out a χ^2 test to test the following null hypothesis:

H_0: No difference in the chances of diarrhoea in patients taking ceftaroline and ceftriaxone.

In Unit 10, you learned about t-tests. In the following activity you will analyse some data from a clinical trial using a t-test.

Computer activity 97 *Does stretching improve movement of a hip joint?*

Ankylosing spondylitis is a form of arthritis that can affect the hip joint. A clinical trial was conducted at the Royal National Hospital for Rheumatic Diseases, in Bath, to see if some extra stretching exercises could be of benefit to such patients. The trial split subjects randomly into two groups: those given a standard treatment (the control group) and those given some stretching exercises as well as the standard treatment (the treatment group).

As part of this trial, the amount of lateral rotation (in degrees) in subjects' right hips was recorded before and after some treatment. Hence the change in lateral rotation could be worked out for every subject.

(Source: Chatfield, C. (1988) *Problem Solving: A Statistician's Guide*, Chapman & Hall.)

(a) What type of trial is this: crossover, matched-pairs or group-comparative? Justify your answer.

(b) What type of data are the 'changes in lateral rotation'?

(c) Given your answers to parts (a) and (b), which type of t-test is likely to be most appropriate for analysing the data: two-sample, one-sample or matched-pairs? (Assume here that any relevant population distribution is normal.)

(d) The null and alternative hypotheses are

$$H_0\colon\ \mu_c = \mu_t$$
$$H_1\colon\ \mu_c \neq \mu_t,$$

where μ_c is the population mean change in lateral rotation for patients given standard treatment and μ_t is the population mean change in lateral rotation for patients given stretching exercises and the standard treatment. Does this correspond to a two-sided or one-sided test?

(e) The data are given in the file **hips.mtw**. Use Minitab to test the hypotheses given in part (d). What are your conclusions about the effect of the stretching exercises?

(f) Give a 95% confidence interval for $\mu_t - \mu_c$. Why does this interval agree with the conclusion you came to in part (e)?

(g) When carrying out parts (e) and (f), you will have used a pooled standard deviation. Was this reasonable? Why or why not?

Summary of Chapter 11

In this chapter, you have learned how to implement different randomisation schemes that come up in clinical trials: systematic allocation, simple randomisation, random assignment, and random assignment with fixed totals. You have also seen how the results from clinical trials can be analysed by using the χ^2 test for contingency tables and by using a t-test.

12 Binomial distribution and two-sample tests

This chapter, which is associated with Unit 12, extends a couple of techniques in Minitab that you have already met. In Subsection 12.1, you will use Minitab to explore the shape of the binomial distribution for different sample sizes and success probabilities. Then, in Subsection 12.2, you will learn how to perform two-sample (unpaired) t-tests that do not rely on the assumption that the samples come from populations whose variances are equal.

12.1 More on the binomial distribution

In Subsection 6.1, you learned how to use Minitab to calculate probabilities from a binomial distribution. However, Subsection 6.1 only dealt with binomial distributions where the probability of success (p) is 0.5. Here you will learn how to obtain probabilities from binomial distributions whatever the value of p. (Note that because p is a probability, only values between 0 and 1 make sense.)

One purpose of this subsection is to compare the results from the different computer activities. Hence it is best to do all the computer activities in this subsection in a single session.

Computer activity 98 *Binomial distribution with $p = 0.7$ and $n = 10$*

Suppose 'DontPanic' is a small driving school and that, of the people whom they teach, 70% pass the driving test at the first attempt. Further suppose that 10 of DontPanic's pupils are due to take the driving test for the first time next month. The number of these pupils who pass next month, x, follows a binomial distribution with $p = 0.7$ and $n = 10$. Obtain this probability distribution, in a column called `prob1`, by doing the following.

- Open a new worksheet in Minitab. (If you have not just started a new session, create a new worksheet via **File** > **New**.)
- In the worksheet, create a column called `x` with the entries `0`, `1`, ..., `10`, to represent the values that x can take (**Calc** > **Make Patterned Data** > **Simple Set of Numbers**). Then, in the **Simple Set of Numbers** dialogue box, put `x` in the **Store patterned data in** field, `0` in the **From first value** field, and `10` in the **To last value** field, and make sure that the **In steps of**, **Number of times to list each value** and **Number of times to list the sequence** fields all contain `1`.)

- Click on **Calc**, then **Probability Distributions** and then **Binomial...**.
- In the **Binomial Distribution** dialogue box, select **Probability** as we wish to calculate probabilities.
- We are interested in a sample of 10 pupils, and the probability that a DontPanic pupil passes first time is 0.7. So enter `10` in the **Number of trials** field and enter `0.7` in the **Event probability** field.
- Fill in the rest of the dialogue box as you would for a binomial distribution when $p = 0.5$ and when the probabilities are to be stored in a column called `prob1` (see Computer activity 53, Subsection 6.1). That is, enter `x` in the **Input column** field and type `prob1` in the **Optional storage** field.

  The dialogue box should now look like Figure 47.

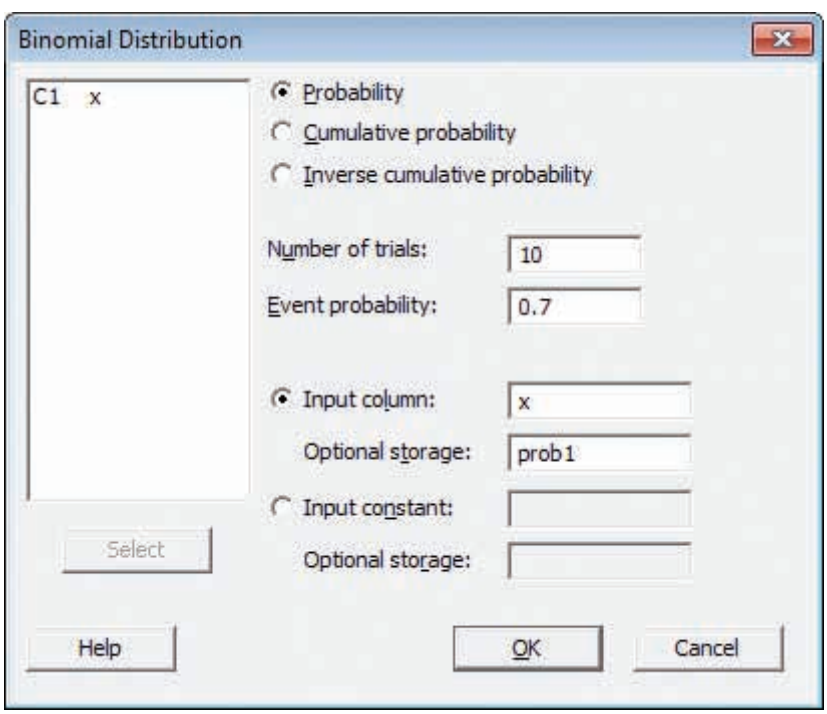


Figure 47 The **Binomial Distribution** dialogue box

- Click on **OK**.

A new column of numbers appears in the worksheet. These numbers are the values of $P(x = 0)$, $P(x = 1)$, ..., $P(x = 10)$. As an example, you should find that 0.266 828 is the probability that exactly 7 of the 10 pupils pass first time.

You will need the results in this worksheet for Computer activity 99. So, if possible, carry straight on to Computer activity 99.

Computer activity 99 *Picturing a binomial distribution*

(a) In Computer activity 54 (Subsection 6.1) you produced a bar chart of a binomial distribution when the probability of success is 0.5. By doing the same as you did in that computer activity, produce a bar chart of the binomial probability distribution you obtained in Computer activity 98 (**Graph** > **Bar Chart**). That is, produce a bar chart of the binomial distribution with $p = 0.7$ and $n = 10$.

(b) Describe the shape of the binomial distribution with $p = 0.7$ and $n = 10$: how many modes does it have, and is it symmetric, left-skew or right-skew?

Computer activity 100 *Binomial distribution with $p = 0.9$ and $n = 10$*

Suppose 90% of the people who take driving lessons from DontPanic eventually pass the driving test in three or fewer attempts. Consider a random sample of 10 people taught by DontPanic who take their test for the first time in a particular month. Now let x denote the number of these people who eventually pass the driving test in three or fewer attempts.

(a) Obtain the probabilities for the number of people taught by DontPanic who pass the driving test in three or fewer attempts. That is, obtain the binomial distribution with $p = 0.9$ and $n = 10$. Store these probabilities in a column labelled `prob3` and then display them as a bar chart.

(b) Compare the shape of this binomial distribution ($p = 0.9$ and $n = 10$) with the binomial distribution you plotted in Computer activity 99 ($p = 0.7$ and $n = 10$).

In Computer activities 99 and 100, you have seen how the shape of the binomial distribution changes when p (but not n) changes. The distribution when $p = 0.9$ (and $n = 10$) is more (left-)skew than the distribution with $p = 0.7$ (and $n = 10$). This is because a binomial distribution is symmetric for $p = 0.5$ and becomes increasingly more skew as the probability of success moves further away from 0.5 (when the value of n does not change). When $p > 0.5$, the probability distribution is left-skew. When $p < 0.5$, the probability distribution is right-skew, and when $p = 0.5$ the probability distribution is symmetric.

What happens when the value of n changes? This is what you will consider in the next activity.

Computer activity 101 *Binomial distribution with $p = 0.7$ and $n = 60$*

Now consider a six-month period, rather than one month. Suppose 60 of DontPanic's pupils take the driving test for the first time in the next six months. The number who will pass, x, follows a binomial distribution with $p = 0.7$ and $n = 60$.

(a) Determine the probabilities of a binomial distribution with $p = 0.7$ and $n = 60$. Store these probabilities in a column labelled `prob6month` and display them as a bar chart.

(b) Comment on the shape of this distribution in comparison to (i) a binomial distribution with $p = 0.7$ and $n = 10$, and (ii) a normal distribution.

In Computer activity 101 you will have found that the binomial distribution with $p = 0.7$ and $n = 60$ is more symmetric than the binomial distribution with $p = 0.7$ and $n = 10$, and that it has a similar shape to the normal distribution. In fact, the binomial distribution becomes steadily more symmetric as the sample size (n) increases, assuming the probability of success (p) remains the same. Further, the binomial distribution has approximately the same shape as a normal distribution when the sample size is large enough.

12.2 More on two-sample tests

In Section 5 of Unit 12, various tests for comparing two population means were reviewed and a new test was introduced. This latter test should be used when there are two samples (which are not matched pairs) from populations that have normal distributions and *we do not want to make the assumption that the variances of the two populations are equal*. We explore this test further in this subsection and compare its results with those obtained when the assumption is made that the two population variances are equal.

It is best to do Computer activities 102 to 104 in a single session. In Computer activity 102 you will draw boxplots for two samples of data and compare their variances. In Computer activity 103 you will then test for equality of the two population means using the new test (no assumption is made that the population variances are equal). Then, in Computer activity 104, you will again test whether the population means are equal, but under the assumption that the population variances are equal. You are then asked to comment on the result of that test in comparison with the result of the test in Computer activity 103.

In Computer activity 105 you perform the same two tests, but for data that are given in the form of summary statistics (as was done for the one-sample z-test in Subsection 7.3).

Computer activity 102 *Multiple boxplots of two samples*

Open the worksheet **Rtimes.mtw** and make sure it is the active window. This worksheet contains data on the time (in seconds) taken for aspirin tablets to release 50% of the painkilling agent ('release times'). The tablets are from two manufacturers, A and B (with 10 from A and 20 from B). The times are given in the worksheet as `ManA` and `ManB`.

(a) Produce a single diagram that contains a boxplot for `ManA` and a boxplot for `ManB`. (The method for getting Minitab to draw such boxplots was described in Computer activity 36 in Subsection 3.3 (**Graph** > **Boxplot**).)

(b) On the basis of the boxplots in part (a):

- Is it reasonable to believe that each sample comes from a population that has a normal distribution?
- Could the two samples have come from populations with the same variance?

(c) Use Minitab to get the variances of `ManA` and `ManB`. (This was covered in Computer activity 32 in Subsection 3.2 (**Stat** > **Basic Statistics** > **Display Descriptive Statistics**).)

Compare the variances you obtain – is it reasonable to assume that they come from populations whose variances are equal?

Suppose that a regulator is interested in whether the times to release 50% of the active agent achieved by tablets from manufacturer A and from manufacturer B are, on average, the same. In other words, the regulator is interested in testing the null hypothesis

$$H_0\colon\ \mu_A = \mu_B$$

against the alternative hypothesis

$$H_1\colon\ \mu_A \neq \mu_B,$$

where μ_A denotes the mean time taken by a tablet produced by manufacturer A to release 50% of the painkilling agent and μ_B denotes the corresponding mean time taken by tablets produced by manufacturer B.

In Computer activity 102 you saw that the sample of tablets from manufacturer A is small (10 tablets) but that it is reasonable to assume that the distributions of release times for tablets from both manufacturers are normal. (Or at least it is not unreasonable to assume this!) This suggests that the two-sample t-test is a reasonable test to use.

However, you also saw in Computer activity 102 that it is *not* reasonable to assume that the variances of the two populations are equal. So a version of the two-sample t-test that does not make this assumption is required. This is what you will use in Computer activity 103.

Computer activity 103 *Two-sample t-test and confidence interval when population variances are not equal*

A two-sample t-test is required that does not make the assumption that population variances are equal based on the sample data in **Rtimes.mtw** to test the null hypothesis

$$H_0\colon\ \mu_A = \mu_B$$

against the alternative hypothesis

$$H_1\colon\ \mu_A \neq \mu_B.$$

The procedure for doing this test in Minitab is very similar to the procedure for doing a two-sample t-test which does make the assumption of equal population variances (Subsection 10.2). So first do the following.

- With **Rtimes.mtw** as the active window, click on **Stat**, then **Basic Statistics** and then **2-Sample t...**.
- In the resulting **Two-Sample t for the Mean** dialogue box, select `Each sample is in its own column` from the drop-down menu and then enter `ManA` in the **Sample 1** field and `ManB` in the **Sample 2** field.

Now, for these data, the assumption that the population variances are equal is not reasonable. So for this two-sample t-test do the following.

- Obtain the **Two-Sample t: Options** dialogue box by clicking on the **Options...** button. In this dialogue box make sure the **Assume equal variances** is *not* selected.

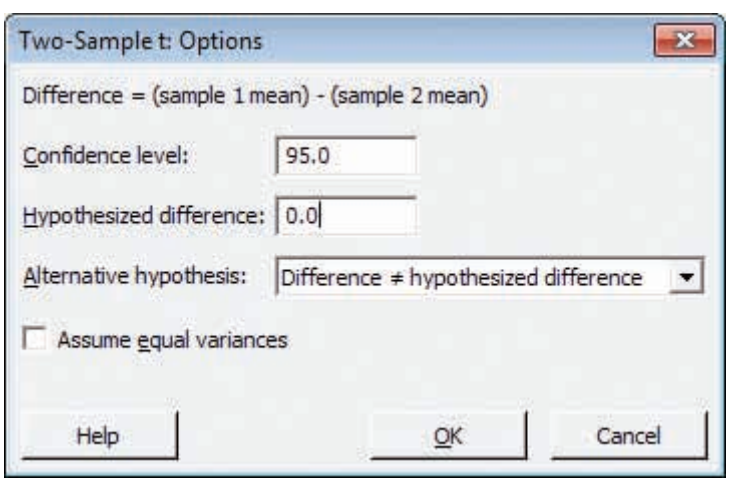


Figure 48 The **Two-Sample t: Options** dialogue box indicating that equal variances are not to be assumed

- Click on **OK** and then on **OK** again.

Using the output given by Minitab, answer the following.

(a) What is the value of the test statistic, and what is the value of the degrees of freedom that Minitab has calculated?

(b) What is the p-value from the test? What conclusion should be drawn from the test?

(c) What is the 95% confidence interval for $\mu_A - \mu_B$, the difference between the two population means? How does this relate to the result of the hypothesis test?

In Computer activity 103 the two sample t-test you performed did not make the assumption that the population variances were equal. But what difference does this assumption make? This is what you will explore in the next computer activity.

Computer activity 104 *Making the (unreasonable) assumption of equal population variances*

Test the same hypotheses as in Computer activity 103, but this time making the assumption that the samples come from two populations whose population variances are equal.

(a) What is the value of the test statistic, and what is the value of the degrees of freedom? How do these compare to the values found in Computer activity 103?

(b) What is the p-value from this test? If the assumptions made were correct, what conclusion would you draw from this test?

(c) Comment on the result of this test in relation to the result found in Computer activity 103.

Comparing the results obtained in Computer activities 103 and 104 illustrates the importance of checking whether population variances should be assumed to be equal. For these data and hypotheses, the conclusions drawn depend on whether or not the population variances are assumed to be equal.

So which set of conclusions is right? Well, in this case, the rule of thumb suggests that the population variances are not equal. So the conclusions drawn from the t-test which does not make this assumption (that is, little evidence that the population means are different) are more defensible.

However, to put this result in perspective, we should mention that often the two forms of the two-sample t-test yield the same conclusion. When this happens, it means that the assumption that the population variances are not equal is not all that important. In such cases, it does not really matter if they are or not!

The following computer activity gives a further comparison of the two forms of the two-sample t-test. Furthermore, in this activity you will perform the two-sample t-test by entering summary statistics to perform the test. Thus, it will be very similar to the way you performed the one-sample z-test in Minitab (Subsection 7.3).

Computer activity 105 *A further comparison of the two t-tests*

Suppose treatments A and B are two treatments for the common cold. Further suppose that summary statistics for the length of time (in days) some patients had symptoms when they were taking either treatment A or treatment B are given in Table 6.

Table 6 Summary statistics for the length of time with symptoms for patients with the common cold

	Sample size	Sample mean	Sample standard deviation
Treatment A	12	6.2	1.5
Treatment B	20	8.1	2.8

As in Computer activities 103 and 104, we want to test the null hypothesis

$$H_0\colon\ \mu_A = \mu_B$$

against the alternative hypothesis

$$H_1\colon\ \mu_A \neq \mu_B,$$

where μ_A and μ_B are the means of the populations from which the samples were taken. This corresponds to testing whether, on average, the two treatments are equally effective.

Using the information in Table 6, it is not possible to decide if it is reasonable that the population distribution for the length of time with symptoms in each treatment group is normal. In such cases, other knowledge about the data has to be used. For example, in this case assume that clinicians believe that for either treatment the population distribution of the length of time with symptoms is normal.

(a) Perform a two-sample t-test that makes no assumption about the population variances being equal. Do this using the following steps.

- Obtain the **Two-Sample t for the Mean** dialogue box (**Stat** > **Basic Statistics** > **2-Sample t**).
- Select `Summarized data` from the drop-down menu.
- Fill in the boxes with the sample size, mean and standard deviation of each sample. The completed dialogue box should look as in Figure 49.

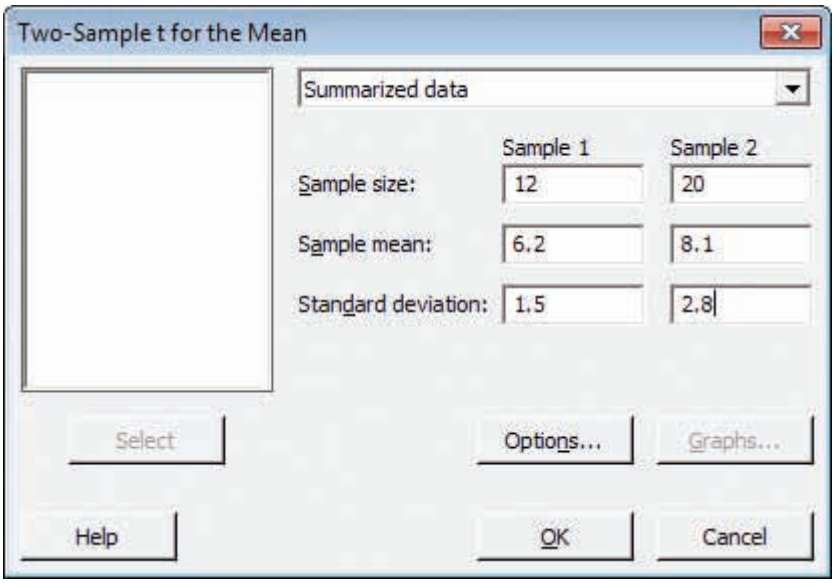


Figure 49 Dialogue box for a two-sample t-test using summary statistics

- Make sure that **Assume equal variances** is *not* selected in the **Two-Sample t: Options** dialogue box.
- Click on **OK** and then **OK** again.

What is the p-value from the test? Assuming its underlying assumptions are satisfied, what conclusion should be drawn from the test?

(b) Perform a two-sample t-test, but this time make the assumption that the population variances are equal.

In other words, obtain the **Two-Sample t for the Mean** dialogue box again (**Stat** > **Basic Statistics** > **2-Sample t**). You should find that all of the fields are filled in for you. However, in the **Two-Sample t: Options** dialogue box, this time make sure that **Assume equal variances** *is* selected.

What is the p-value from this test? Assuming its underlying assumptions are satisfied, what conclusion should be drawn from the test?

(c) Compare the results of the two t-tests you have performed.

(d) Which of the two t-tests was the correct test to use for these data and hypotheses?

Summary of Chapter 12

In this chapter, you have extended your knowledge about two different tasks in Minitab: calculating binomial probabilities and performing a two-sample t-test.

You have learned how to calculate binomial probabilities whatever the value of the success probability, p, is assumed to be. You have seen that when p moves away from $p = 0.5$ and gets closer to $p = 1$, the probability distribution gets more skew. You have also seen that when the number of trials increases, the probability distribution becomes more symmetric.

For the two-sample t-test, you have extended the range of data to which it can be applied using Minitab. You have learned how to perform a version of this test that does not rely on the assumption of equal population variances. You have seen that whether or not this assumption is made can make a difference to the conclusions that are drawn from the test. You have also learned how to perform the test when the data are given in summarised form.

Minitab quick reference guide

The following list summarises the Minitab terms and commands used in M140.

- Active (data) window: The data worksheet that Minitab will use. It is denoted by *** after the name of the worksheet. (Computer activity 2, A guided tour of Minitab)
- Bar chart: **Graph** > **Bar Chart**. (Computer activity 54, Subsection 6.1)
- Binomial probabilities (calculating): **Calc** > **Probability Distributions** > **Binomial**. (Computer activity 53 and Computer activity 55, Subsection 6.1; Computer activity 98, Subsection 12.1)
- Boxplot (Multiple): **Graph** > **Boxplot** > **Multiple Y's, Simple**. (Computer activity 36, Subsection 3.3)
- Boxplot (Simple): **Graph** > **Boxplot** > **One Y, Simple**. (Computer activity 34, Subsection 3.3)
- Calculating variables: **Calc** > **Calculator**. (Computer activity 6, Subsection 1.2)
- Chi-square test: **Stat** > **Tables** > **Chi-Square Test for Association**. (Computer activity 68, Subsection 8.1)
- Column of numbers (generate): **Calc** > **Make Patterned Data** > **Simple Set of Numbers**. (Computer activity 43, Subsection 4.2)
- Confidence interval (based on one-sample z-test): **Stat** > **Basic Statistics** > **1-Sample Z**. (Computer activity 75, Subsection 9.2)
- Confidence interval for the mean response: First find the equation of the least squares regression line using Minitab (**Stat** > **Regression** > **Regression** > **Fit Regression Model**). Then **Stat** > **Regression** > **Regression** > **Predict**. (Computer activity 82, Subsection 9.4)
- Contingency table (creating): **Stat** > **Tables** > **Cross Tabulation and Chi-Square**. (Computer activity 93, Subsection 11.1)
- Contingency table (entering in Minitab): **File** > **New**. Enter the row categories in the first column and the column categories as variable names. The marginal totals are not entered in the worksheet. (Computer activity 71, Subsection 8.2)
- Correlation: **Stat** > **Basic Statistics** > **Correlation**. (Computer activity 73, Subsection 9.1)
- Data window: A window containing a dataset. (Computer activity 1, A guided tour of Minitab)
- Entering consecutive integers: **Calc** > **Make Patterned Data** > **Simple Set of Numbers**. (Computer activity 43, Subsection 4.2)
- Exiting Minitab: **File** > **Exit**. (Computer activity 1, A guided tour of Minitab)

- Finding windows: Click on the name of the window in the **Window** menu. (Computer activity 3, A guided tour of Minitab)
- Histogram: **Graph** > **Histogram**. (Computer activity 18, Subsection 1.5)
- Interquartile range: **Stat** > **Basic Statistics** > **Display Descriptive Statistics** and click on **Statistics**. (Computer activity 32, Subsection 3.2)
- Least squares regression line (adding to scatterplot): **Graph** > **Scatterplot**. Make the Graph window containing the scatterplot active. Then **Editor** > **Add** > **Regression Fit**. (Computer activity 49, Subsection 5.2)
- Least squares regression line (calculating): **Stat** > **Regression** > **Regression** > **Fit Regression Model**. (Computer activity 48, Subsection 5.2)
- Maximum: **Stat** > **Basic Statistics** > **Display Descriptive Statistics**. (Computer activity 31, Subsection 3.2)
- Mean: **Stat** > **Basic Statistics** > **Display Descriptive Statistics**. (Computer activity 31, Subsection 3.2)
- Median: **Stat** > **Basic Statistics** > **Display Descriptive Statistics**. (Computer activity 31, Subsection 3.2)
- Minimum: **Stat** > **Basic Statistics** > **Display Descriptive Statistics**. (Computer activity 31, Subsection 3.2)
- New worksheet: **File** > **New** and select 'Minitab Worksheet'. (Computer activity 40, Subsection 4.2)
- One-sided tests: In the dialogue box for the test, click on **Options...** and then select either `Difference > hypothesized difference` or `Difference < hypothesized difference` in the **Alternative** field of the dialogue box. (Computer activity 90 and Computer activity 91, Subsection 10.4)
- Pasting a graph into a word-processor document: Make the Graph window active then **Edit** > **Copy Graph**. Then in the word-processor document **Edit** > **Paste**. (Computer activity 25, Subsection 1.6)
- Pasting a worksheet into a word-processor document: In Minitab, highlight the cells to be pasted then **Edit** > **Copy Cells**. Then in the word-processor document **Edit** > **Paste**. (Computer activity 23, Subsection 1.6)
- Pasting output from Minitab into a word-processor document: In Minitab, highlight the output in the Session window to be pasted then **Edit** > **Copy**. Then in the word-processor document **Edit** > **Paste**. (Computer activity 24, Subsection 1.6)
- Prediction interval: First find the equation of the least squares regression line using Minitab (**Stat** > **Regression** > **Regression** > **Fit Regression Model**). Then **Stat** > **Regression** > **Regression** > **Predict**. (Computer activity 82, Subsection 9.4)

- Printing graphs: Make sure the Graph window is active then **File** > **Print Graph**. (Computer activity 22, Subsection 1.6)
- Printing output: Make sure the Session window is active, highlight the area to be printed, then **File** > **Print Session Window**. (Computer activity 21, Subsection 1.6)
- Printing worksheets: Make sure the Data window is active then **File** > **Print Worksheet**. (Computer activity 20, Subsection 1.6)
- Quartiles: **Stat** > **Basic Statistics** > **Display Descriptive Statistics**. (Computer activity 31, Subsection 3.2)
- Random allocation: **Calc** > **Random Data** > **Binomial**. (Computer activity 94, Subsection 11.1)
- Random allocation (keeping totals fixed): Use systematic allocation and then **Calc** > **Random Data** > **Sample from Columns**. (Computer activity 95, Subsection 11.1)
- Random numbers: **Calc** > **Random Data** > **Integer**. (Computer activity 40, Subsection 4.2)
- Random sample (from column of numbers): **Calc** > **Random Data** > **Sample From Columns**. (Computer activity 44, Subsection 4.2)
- Range: **Stat** > **Basic Statistics** > **Display Descriptive Statistics** and click on **Statistics**. (Computer activity 32, Subsection 3.2)
- Residual plot: **Stat** > **Regression** > **Regression** > **Fit Regression Model**. Click on **Graphs...**, select `Regular` and enter the explanatory variable in the **Residuals versus the variables** field. (Computer activity 51, Subsection 5.3)
- Save (session): **File** > **Save Project** or **File** > **Save Project As**. (Computer activity 26, Subsection 1.6)
- Save (window): **File** > **Save Graph**, **File** > **Save Graph As**, **File** > **Save Session**, **File** > **Save Session As**, **File** > **Save Current Worksheet** or **File** > **Save Current Worksheet As**. (Computer activity 27, Subsection 1.6)
- Scatterplot: **Graph** > **Scatterplot** and select **Simple**. (Computer activity 4, Subsection 1.1)
- Session window: The window in which results (apart from graphs) produced by Minitab are displayed. (Computer activity 1, A guided tour of Minitab)
- Setting the seed (of the random number generator): **Calc** > **Set base**. (Computer activity 92, Subsection 11.1)
- Sign test: **Stat** > **Nonparametrics** > **1-Sample Sign**. (Computer activity 57, Subsection 6.2)
- Standard deviation: **Stat** > **Basic Statistics** > **Display Descriptive Statistics**. (Computer activity 31, Subsection 3.2)
- Starting Minitab: Click on the **Minitab 17** icon on your desktop, or select 'Minitab 17 Statistical Software' from your list of programs (Computer activity 1, A guided tour of Minitab)

- Stemplot: **Graph** > **Stem-and-Leaf**. (Computer activity 10, Subsection 1.3)
- Systematic allocation: **Calc** > **Make Patterned Data** > **Simple Set of Numbers**. (Computer activity 93, Subsection 11.1)
- Tabulating numbers: **Stat** > **Tables** > **Tally Individual Variables**. (Computer activity 41, Subsection 4.2)
- t-test (matched-pairs): **Stat** > **Basic Statistics** > **Paired t**. (Computer activity 88, Subsection 10.3)
- t-test (one-sample): **Stat** > **Basic Statistics** > **1-Sample t**. (Computer activity 84, Subsection 10.1)
- t-test (two-sample): **Stat** > **Basic Statistics** > **2-Sample t**. If the assumption of equal population variances is to be made, make sure that **Assume equal variances** is selected in the **Two-Sample t: Options** dialogue box. (Computer activity 86, Subsection 10.2; Computer activity 103 and Computer activity 104, Subsection 12.2)
- Variance: **Stat** > **Basic Statistics** > **Display Descriptive Statistics** and click on **Statistics**. (Computer activity 32, Subsection 3.2)
- Worksheet (opening): **File** > **Open Worksheet**. (Computer activity 2, A guided tour of Minitab)
- z-test (one-sample): **Stat** > **Basic Statistics** > **1-Sample Z**. (Computer activity 64, Subsection 7.3)

Solutions to computer activities

Solution to Computer activity 1

(a) **Calculator...** is listed in the **Calc** menu. (You may also have found **Microsoft Calculator** listed in the **Tools** menu.)

(b) **Empirical CDF...** is listed in the **Graph** menu.

(c) **1-Sample Z...** is listed in the **Basic Statistics** menu, which is a submenu of **Stat**.

(d) The types of commands available in the menus may be broadly summarised as follows.

- The **File** and **Edit** menus contain commands for opening, handling and editing files.
- The **Data** menu allows you to manipulate data in the Data window.
- Statistical calculations may be carried out using the **Calc** and **Stat** menus.
- The **Graph** menu is used to create graphs and diagrams.
- The **Editor** and **Tools** menus allow you to customise aspects of Minitab.
- The **Window** menu is for activating or rearranging windows.
- The **Help** and **Assistant** menus provide access to help and online resources. Some of the information is about using Minitab and some about statistics in general – however, much of this is beyond the scope of this module.

Solution to Computer activity 2

The data in the worksheet are laid out in a tabular format. The dataset itself is laid out in the white cells, with the grey cells at the top and side used for labelling information.

Each row represents a different observation or *case*. Here, each case corresponds to information about a particular year.

Each column represents a different *variable*, that is, a different sort of measurement made about the cases.

In this worksheet there are four variables: the calendar year, the number of new large marine species discovered that year, the total number of large marine species known about in that year and the total number of large marine species predicted by a statistical model. The two grey rows at the top give generic names for the variables, `C1`, `C2`, `C3` and `C4`, and descriptive names, `year`, `new`, `total` and `model`. Generally, either the generic name or the descriptive name can be used to specify a particular variable. However, it is good practice to use the descriptive name where possible as it reduces the chance of specifying the wrong variable by accident.

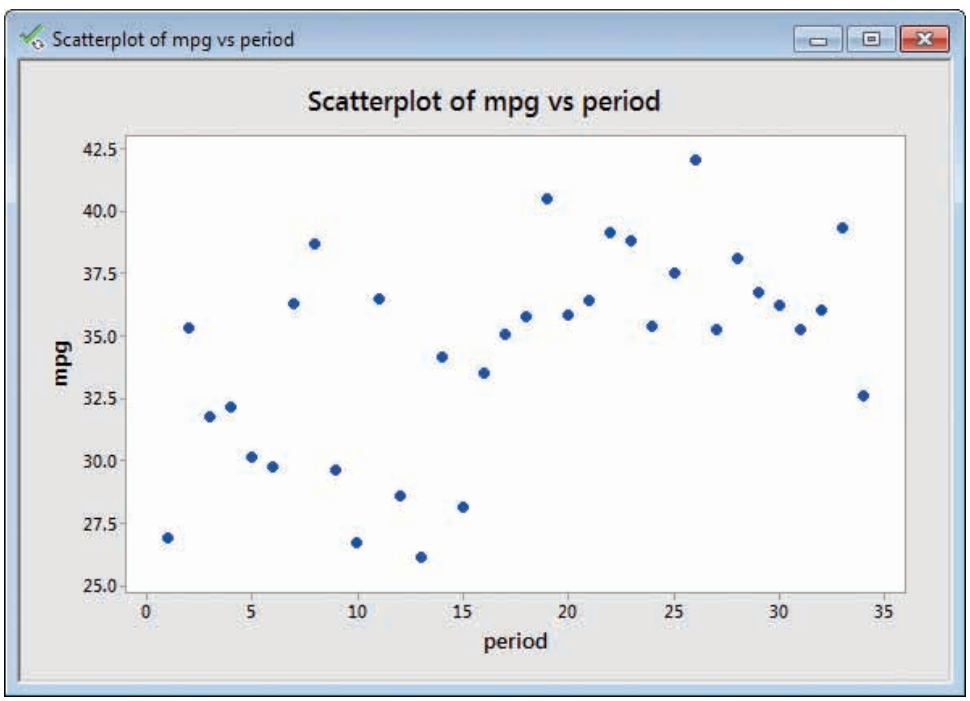
Descriptive names are usually chosen to be short but meaningful. Abbreviations are often used.

Solution to Computer activity 3

The worksheet **mpg.mtw** is opened in the same way that the worksheet **bigfish.mtw** was opened in Computer activity 2. The only difference is that the file **mpg.mtw** should be selected instead of **bigfish.mtw** in the **Open Worksheet** dialogue box.

Solution to Computer activity 5

(a) The Graph window with the required scatterplot is given below.



This is obtained by doing the following.

- Choose **Scatterplot...** from the **Graph** menu, and select **Simple**.
- In the **Scatterplot: Simple** dialogue box, copy `mpg` to the **Y variables** field and `stop` to the **X variables** field, and click on **OK**.

(b) The scatterplot shows a vague upward trend from the lower left-hand corner to the upper right-hand corner, suggesting that the petrol consumption, in terms of miles per gallon, has increased over time.

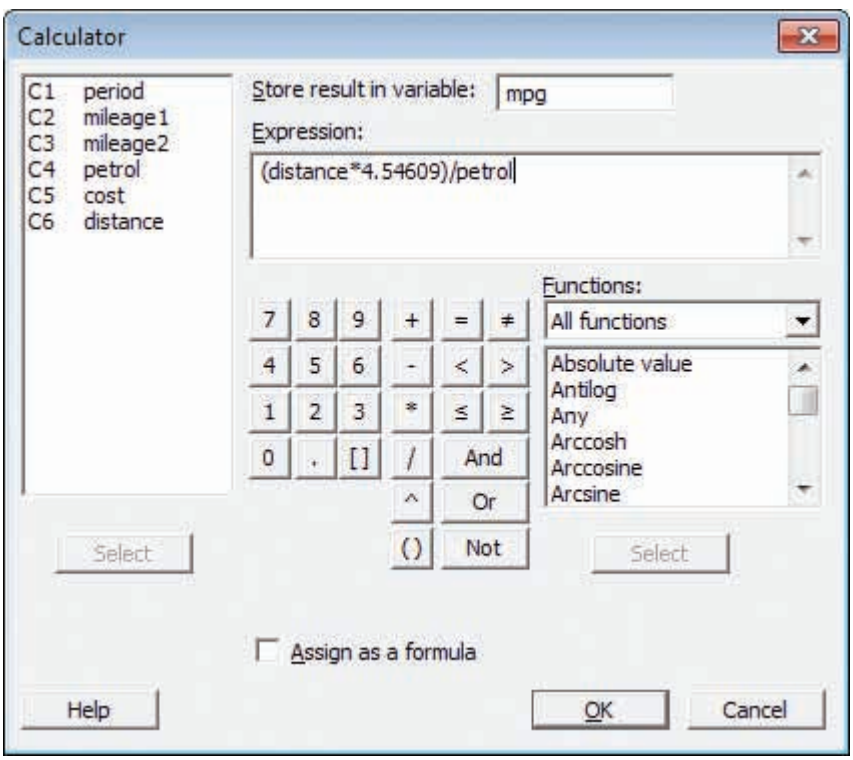
Solution to Computer activity 7

In Minitab, the **Calculator** dialogue box is required. This is obtained by clicking on **Calc** and then **Calculator...**.

The name for the new variable is `mpg`. This name should be entered in the **Store result in variable** field. The equation that needs to be entered in the **Expression** field of the **Calculator** dialogue box is

```
(distance*4.54609)/petrol
```

The **Calculator** dialogue box should therefore look as follows.

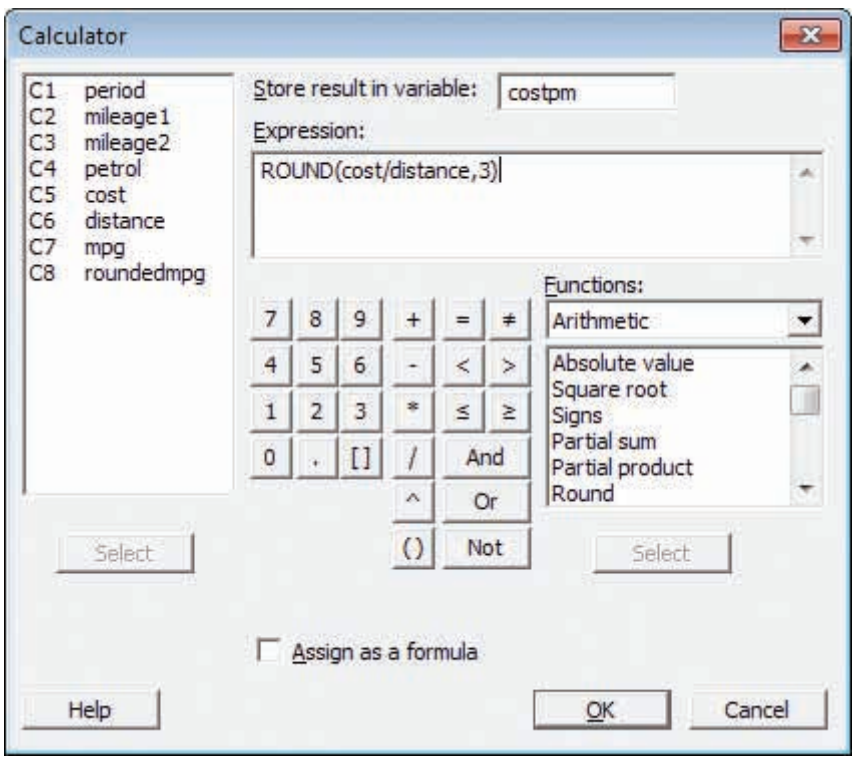


The value of `mpg` corresponding to the first stop is 26.8844.

Solution to Computer activity 9

Start by opening the **Calculator** dialogue box (**Calc** > **Calculator**).

The variable name `costpm` should be entered in the **Store result in variable** field, and `ROUND(cost/distance,3)` should be entered in the **Expression** field. So the completed **Calculator** dialogue box should look as follows.



The value of `costpm` corresponding to the first stop is 0.188 (in £ per mile).

Solution to Computer activity 10

The batch size is 18. So the median is the average of the 9th and 10th values. From the `Count` column you can see that there are 8 values at or below level 6. So the two leaves on level 7 must be the 9th and 10th data values. Thus the 9th and 10th data values are 7500 thousand tonnes and 7700 thousand tonnes. As $(7500 + 7700)/2 = 7600$, the median value is 7600 thousand tonnes.

Solution to Computer activity 11

(a) The following stemplot appears in the Session window. It is a stemplot of all the distances given in the spreadsheet.

```
Stem-and-leaf of distance  N  = 15
Leaf Unit = 0.10


 1    11  9
 4    12  357
 6    13  89
 7    14  3
(3)   15  124
 5    16  2447
 1    17
 1    18  0
```

The batch size is 15, so the median is the 8th data value. You can immediately see that the median corresponds to a leaf on level 15, as the corresponding value in the `Count` column is in brackets.

There are 7 data values up to and including the leaf on level 14, so the median corresponds to the first leaf on level 15 – the leaf '1'. The leaf units are 0.1, so the median is 15.1 metres.

(b) Two new stemplots now appear in the Session window: a stemplot for the values corresponding to `pool = 1` and another stemplot for the values corresponding to `pool = 2`.

```
Stem-and-leaf of distance  pool = 1    N  = 7
Leaf Unit = 0.10


 2    12  57
 2    13
 2    14
 3    15  4
(3)   16  247
 1    17
 1    18  0


Stem-and-leaf of distance  pool = 2    N  = 8
Leaf Unit = 0.10


 1  11  9
 2  12  3
 4  13  89
 4  14  3
 3  15  12
 1  16  4
```

(c) In the first pool there were 7 senior male athletes, so the median is the 4th largest value. From the `Count` column it is clear that this value lies on level 16. As there are 3 values up to and including the last leaf on level 15, the 4th largest value is the first leaf on level 16. This corresponds to a best throw of 16.2 metres.

In the second stemplot none of the values in the `Count` column has brackets round it, so the median would be on one of the levels with the highest value – level 13 and level 14 in this case. It is the average of the last leaf on level 13 and the first leaf on level 14: corresponding to best throws of 13.9 metres and 14.3 metres, respectively. Hence, as $(13.9 + 14.3)/2 = 14.1$, the median best throw in pool 2 is 14.1 metres.

Solution to Computer activity 12

(a) The stemplot is obtained by entering `time` in the **Graph variables** field of the **Stem-and-Leaf** dialogue box (**Graph** > **Stem-and-Leaf**). In the dialogue box the **By variable** field should have been left blank.

The stemplot produced by Minitab is as follows.

```
Stem-and-leaf of time   N   = 31
Leaf Unit = 0.10

 2    47   78
 8    48   022789
 12   49   0466
 13   50   8
 14   51   1
 14   52
 15   53   9
 15   54
(2)   55   13
 14   56   66
 12   57   22348
 7    58   399
 4    59   24
 2    60   12
```

The median is a value on level 55. Since there are 31 values, the median is the 16th largest value, corresponding to the first leaf on level 55: so the median is 55.1 seconds.

(b) The separate stemplots are obtained in the same way as in part (a), but with the additional step of entering the variable `sex` in the **By variable** field. The two stemplots are as follows.

```
Stem-and-leaf of time   sex = 1     N   = 15
Leaf Unit = 0.10

 1    53   9
 1    54
 2    55   1
 4    56   66
(5)   57   22348
 6    58   39
 4    59   24
 2    60   12
```

```
Stem-and-leaf of time   sex = 2     N   = 16
Leaf Unit = 0.10

 2   47   78
 8   48   022789
 8   49   0466
 4   50   8
 3   51   1
 2   52
 2   53
 2   54
 2   55   3
 1   56
 1   57
 1   58   9
```

The median for the women is the 8th largest value, which is 57.4 seconds. (This is a value on level 57 as indicated by the brackets in the `Count` column. It is the 4th leaf on this level as there are 4 values up to and including the preceding level, level 56.)

The median for the men is the average of the 8th and 9th largest values. Thus the median is $(48.9 + 49.0)/2 = 48.95$, which rounds to 49.0 seconds. (This is calculated using the values corresponding to the last leaf on level 48 and the first leaf of level 49, the two levels with the highest value in the `Count` column.)

Solution to Computer activity 13

The stemplot produced by Minitab is as follows.

```
Stem-and-leaf of price  N  = 26
Leaf Unit = 10


 3    1  789
 10   2  0223333
(6)   2  566779
 10   3  00124
 5    3  5
 4    4  02
 2    4
 2    5
 2    5
 2    6  4
 1    6  9
```

Looking at the stem of the stemplot, all but one of the numbers is repeated. Also, for each row of the stemplot, the leaves are in the range 0 to 4 or in the range 5 to 9. Thus, this is a stretched stemplot. Furthermore, this is the same amount of stretching that was identified in Activity 17(d) of Unit 1 (Subsection 5.1) as 'just right'.

Solution to Computer activity 15

(a) The stemplot produced by Minitab is as follows.

```
Stem-and-leaf of price  N  = 26
Leaf Unit = 10


 1    1  7
 3    1  89
 4    2  0
 10   2  223333
 11   2  5
(4)   2  6677
 11   2  9
 10   3  001
 7    3  2
 6    3  45
 4    3
 4    3
 4    4  0
 3    4  2


HI 64, 69
```

(b) Minitab has identified the televisions costing £649 and £699 as outliers. These are the same two televisions that were picked out as outliers in Activity 17 of Unit 1 (Subsection 5.1).

(c) Here, with the outliers trimmed, Minitab uses a leaf unit of 10 and splits each level into five parts. In Computer activity 13, without the outliers trimmed, the leaf unit was still 10 but each level was split into only two parts. So the amount of squeezing/stretching used automatically by Minitab depends on whether or not the outliers are trimmed.

Solution to Computer activity 16

(a) The default stemplot is obtained from the **Stem-and-Leaf** dialogue box (**Graph** > **Stem-and-Leaf**), with `salary` entered in the **Graph variables** field and all the other fields left blank. It is as follows.

```
Stem-and-leaf of salary  N  = 20
Leaf Unit = 1000


 9    4   445555778
(6)   5   123689
 5    6   0
 4    7   48
 2    8   2
 1    9
 1    10
 1    11  4
```

It looks like there is a clear outlier, with value 114. The leaf unit is 1000 and each level has only one part.

(b) Ticking the **Trim outliers** box in the **Stem-and-Leaf** dialogue box (and leaving the **Increment** field blank) results in the following stemplot.

```
Stem-and-leaf of salary  N  = 20
Leaf Unit = 1000


 2    4   44
 9    4   5555778
(3)   5   123
 8    5   689
 5    6   0
 4    6
 4    7   4
 3    7   8


HI 82, 114
```

This stemplot has been stretched, with each level split into two parts. Two outliers have been trimmed: 82 and 114.

(c) If each level has only one part, then the separation between one level and the next is equal to ten leaf units. So the correct increment to use is ten times the leaf unit, namely 10 000. However, any values between 5001 and 10 000 will work. The value 10 001 will produce a stemplot with just two levels.

(d) If each level is split into five parts, then the separation between one part and the next is equal to two leaf units. So the correct increment to use is 2000. However, any value between 1001 and 2000 will work. Values of 1000 or less will produce a very long stemplot with a different leaf unit, and the value 2001 will produce the same stemplot as in part (b).

Solution to Computer activity 17

This bar, according to the horizontal scale, corresponds to coal production figures between 8000 and 10 000 (in thousand tonnes). Its height on the vertical scale is 6, so there are six such regions. This matches the stemplot as the third row from the bottom of the stemplot has six leaves – three 8s and three 9s (which mean 8000 and 9000 (thousand tonnes)).

Solution to Computer activity 19

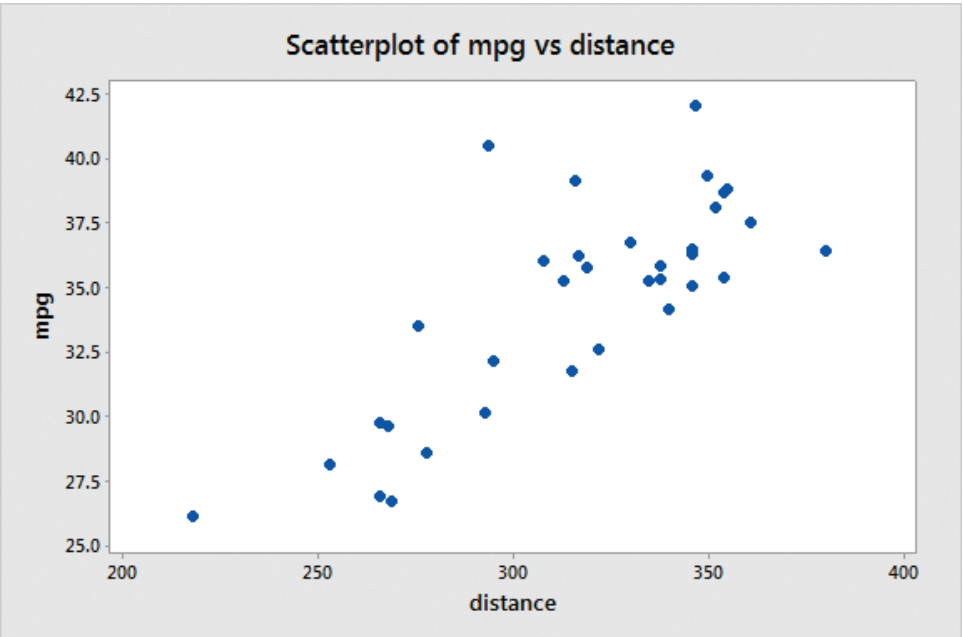
(a) The stemplot is obtained by entering `distance` in the **Graph variables** field of the **Stem-and-Leaf** dialogue box (**Graph** > **Stem-and-Leaf**) and making sure the other fields are blank.

```
Stem-and-leaf of distance  N  = 34
Leaf Unit = 10


 1   2  1
 1   2
 2   2  5
 8   2  666677
 11  2  999
 17  3  011111
 17  3  23333
 12  3  4444455555
 2   3  6
 1   3  8
```

Assuming that the peak at level 3 and leaves '0' and '1' is not significant, the stemplot is bimodal with one mode at level 2 (and leaves '6' or '7') and a bigger mode at level 3 (and leaves '4' and '5'). The distribution appears to be left-skew as the lower values of the distribution appear to be more spread out than the upper values.

(b) The scatterplot is obtained by copying `mpg` to the **Y variables** field of the **Scatterplot - Simple** dialogue box (**Graph** > **Scatterplot** > **Simple**) and copying `distance` to the **X variables** field.

The scatterplot shows an upward trend. So it appears that the further the car travelled between stops, the higher the miles per gallon.



Solution to Computer activity 28

(a) The mean is bigger than the median as its line is further to the right. The difference between the median and the mean is about £10.

(b) The median does not change when the price of the most expensive television is increased. It does not matter how expensive this television becomes, the median stays the same.

The mean increases as the price of the most expensive television increases. This means that as this most expensive television gets closer to £400, the mean gets further away from the median. If the price is allowed to become even bigger, then the mean will continue to move away from the median.

(c) The median still does not change as the price of the second most television increases. The mean does increase as the price of the second most expensive television increases. When both the most expensive and the second most expensive television cost £400, the mean is about £25 more than the median.

(d) The data consist of the prices of twenty televisions. So the median is based on the 10th and 11th most expensive televisions. This means that the prices of the nine most expensive televisions can be increased without changing the median.

Solution to Computer activity 29

(a) The position of the fulcrum represents the mean of the combined batch (that is, for batch A combined with batch B).

(b) When the fulcrum is at 161.9 the bar balances, confirming that 161.9 is indeed the mean of the combined batch.

(c) When the weights on both sides are doubled, the position of the fulcrum that balances the bar does not change; nor does it change when the weights on both sides are tripled. This is to be expected, as Rule 1 for weighted means states that the weighted mean depends on the relative sizes of the weights. So multiplying both weights by the same amount does not change the weighted mean (balance point).

(d) You will have found the balance point is between the two batch means. Also you will have found that the weighted mean is closer to the batch mean with the higher weight. This is to be expected from Rule 2 for weighted means, which states that the weighted mean of two numbers always lies between the numbers and is nearer the number that has the larger weight.

(e) The balance point is at 152, halfway between 119 and 185, the two batch means. It does not change so long as the weights are equal. This is expected by Rule 3 for weighted means, which states that if the weights are equal, then the weighted mean of two numbers is the number halfway between them.

Solution to Computer activity 30

(a) The range is £180, the interquartile range is £50 and the standard deviation is £47.

(b) When the price of the most expensive small television increases, both the range and standard deviation increase. The interquartile range does not change.

(c) When the price of the cheapest small television falls, the range and standard deviation increase. The interquartile range does not change.

(d) When the price of an average-priced small television changes, the standard deviation changes. The range does not change so long as the price of such a television does not become lower than the cheapest television or higher than the most expensive television. The interquartile range only changes if the price becomes lower than £130 (the first quartile for the original data) or higher than £180 (the third quartile for the original data), and then only by a limited amount.

(e) The interquartile range is a resistant measure. It does not change much, if at all, when individual data values are changed.

The standard deviation is a sensitive measure. Changing any data value changes the value of the standard deviation.

The range is also a sensitive measure. Although changing the value of data points that are not extreme does not change the range, the range is affected if an extreme data point (that is, the lowest or highest) is changed by even just a small amount.

Solution to Computer activity 31

(a) The mean, minimum, maximum and median are clearly labelled. You probably guessed that `N` is the batch size and `StDev` means the standard deviation, and maybe you also realised that `Q1` and `Q3` are what are called the lower quartile and the upper quartile in the unit texts. Minitab refers to these as the *first quartile* and the *third quartile*. (The median is the second quartile in this terminology. You have already seen that the median has other names, such as the 50th percentile.)

This leaves the quantities labelled `N*` and `SE Mean`. `N*` is the number of missing data values – in this batch, as in most of those you will meet in M140, there are no missing values, so it is 0. `SE Mean` relates to a quantity that you will learn about later, in Unit 7.

(b) The range is `Maximum` $-$ `Minimum` $= 270 - 90 = 180$.

The interquartile range is (in Minitab's notation)
`Q3` $-$ `Q1` $= 180 - 130 = 50$.

Finally, the variance is $(\texttt{StDev})^2 = (47.0)^2 = 2209.0$.

Solution to Computer activity 32

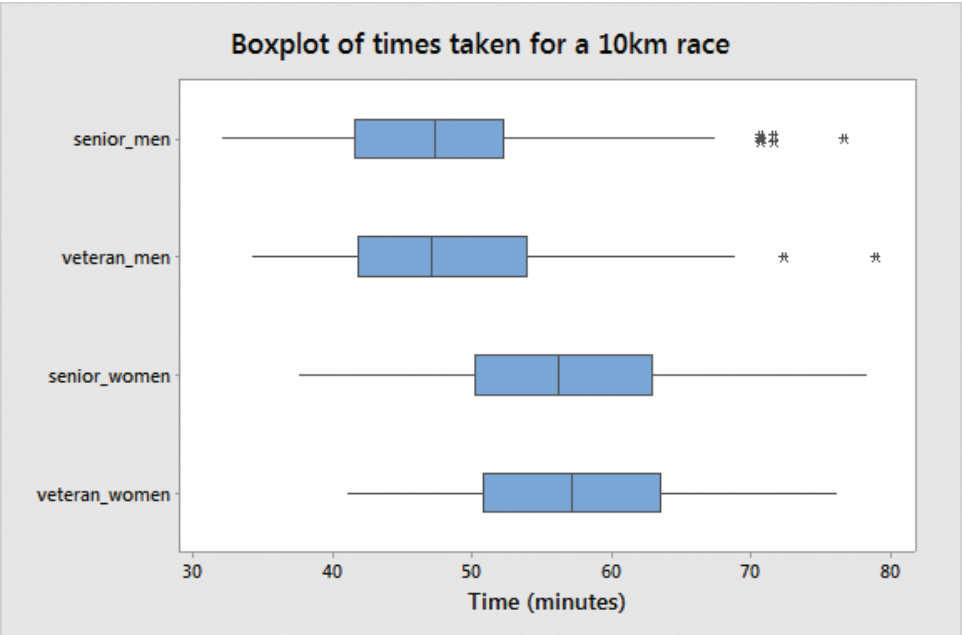
You should see (in the Session window) that the range and interquartile range (which Minitab labels as `IQR`) do indeed take the same values as in Computer activity 31. The variance is given as 2206.3, which looks rather different from the 2209.0 that you should have calculated in Computer activity 31 – but the difference is due to the fact that the standard deviation displayed by Minitab is rounded. Using an unrounded value of the standard deviation would have led to the result for the variance that Minitab displays.

Solution to Computer activity 37

(a) The boxplots are created by selecting **Multiple Y's, Simple** on the **Boxplots** dialogue box (**Graph** > **Boxplot** > **Multiple Y's, Simple**). In the **Boxplot: Multiple Y's, Simple** dialogue box, `senior_men`, `veteran_men`, `senior_women` and `veteran_women` should be placed in the **Graph variables** field. Selecting **Transpose value and category scales** in the **Boxplot: Scale** dialogue box ensures that horizontal boxplots are displayed.

(b) The title and label for the axis can be changed by altering the entries in the **Text** fields of the **Edit Title** and **Edit Axis Label** dialogue boxes. These dialogue boxes are obtained by double-clicking on the title and the axis label for the boxplot, or by using **Editor** > **Select item** and **Editor** > **Edit**.

The boxplots with a suitable title and axis label are given below.



(c) The boxplots for the two groups of women's times are shifted to the right compared with the boxplots for the men's times. So the women generally took longer to complete the race than the men. However, there is some overlap, which means that some of the women did beat some of the men.

(d) The boxplot for the veteran men is similar to the boxplot for the senior men. Also, the boxplot for the veteran women is similar to the boxplot for the senior women. So it appears that age made little difference to the distribution of times for the men and women.

Solution to Computer activity 38

(a) When a member of the module team did this, the sample they got was 1, 4 and 4. So the median response they obtained was 4. This is different to the population median response. However, the median for your sample may have been the same.

(b) When a member of the module team did this, two of the 10 samples had a median of 3. It is likely that you found up to six of your samples had a median of 3.

(c) The table completed by a member of the module team is as follows. Yours is likely to be different, though the numbers should be similar.

Sample median response	Number of samples	Proportion of samples
1	27	0.25
2	18	0.16
3	31	0.28
4	30	0.27
5	4	0.04
Total	110	1.00

(d) This compares reasonably well with Table 10 in Unit 4. Most of the sample median responses were 1, 2, 3 or 4, and very few of them were 5.

Solution to Computer activity 39

(a) When a member of the module team did this, 30 of the 100 samples of size 3 had a median of 3. However you may have found a different number of samples had a median of 3.

(b) When the sample size was 5, there were 31 samples with a median of 3.

(c) When the sample size was 11, there were 50 samples with a median of 3.

(d) The table completed with the results from a member of the module team was as follows.

Sample size	Number of samples with sample median 3
3	30
5	31
11	50
21	68
41	80
81	98

This suggests that if more than 90% are required to have a sample median of 3, a sample size of 81 is required.

Solution to Computer activity 40

(a) When a member of the module team completed this activity, the following random numbers were generated.

```
 6  98  98  43  27  40  49  36  67  50
54  53  83   1  28   6  43  49  36  50
11   9  41  12  73  43  75  59  62  46
```

The values you obtained are unlikely to be exactly the same as these!

(b) Reading along the rows in the list of random numbers given in the solution to part (a), the following numbers are repeated.

```
98  6  43  49  36  50
```

The random numbers that are repeated in your sample are likely to be different. You may even have found that none of the random numbers in your sample were repeated.

(c) The rule 'take the numbers in the order in which they are generated, ignoring repeats' means that the sample of 15 corresponds to the first 15 unique random numbers. Using the random numbers given in the solution to part (a), this corresponds to the following sample.

```
 6  98  43  27  40  49  36  67  50  54
53  83   1  28  11
```

Solution to Computer activity 41

You should have identified the same repeated values and hence the same labels for the 15 individuals to be selected from the population of 100 that you identified in Computer activity 40.

Solution to Computer activity 42

Open a new worksheet in Minitab by selecting 'Minitab Worksheet' in the **New** dialogue box (**File** > **New**). Then in the **Integer Distribution** dialogue box (**Calc** > **Random Data** > **Integer**) complete the values as follows.

- **Number of rows of data to generate**: `15`
- **Store in column(s)**: `C1`
- **Minimum value**: `1`
- **Maximum value**: `1000`

You may have chosen a number larger than 15 for the number of rows to ensure you have enough random numbers to allow for repeats; that is fine. The first 10 unique numbers in `C1` are labels for a sample of size 10 from a population of size 1000.

For example, when a member of the module team did this, the numbers they obtained were as follows.

```
153  173  645  717  383  888  207  523
469  315   63  551  176  994  744
```

There are no repeated numbers in this list. So the sample is just the first 10 numbers.

```
153  173  645  717  383  888  207  523
469  315
```

Solution to Computer activity 45

(a) Open a new worksheet in Minitab by selecting 'Minitab Worksheet' in the **New** dialogue box (**File** > **New**).

There are then two ways of obtaining a sample of the labels in this new Minitab worksheet.

- **Method 1**

  In the new worksheet you could have created a column containing the numbers 1 to 86. (**Calc** > **Make Patterned Data** > **Simple Set of Numbers** with `population` in the **Store patterned data in** field, `1` in the **From first value** field, `86` in the **To last value** field, and `1` in the other three fields of the resulting dialogue box.) Then, in the **Sample From Columns** dialogue box (**Calc** > **Random Data** > **Sample From Columns**), enter `15` in the **Number of rows to sample** field, copy `population` into the **From Columns** field, and enter `sample` in the **Store samples in** field. The labels for the sampled staff members are then in the variable `sample`.

- **Method 2**

  Alternatively, in the **Integer Distribution** dialogue box (**Calc** > **Random Data** > **Integer**) enter the following: `30` in **Number of rows of data to generate**; `C1` in **Store in column(s)**; `1` in **Minimum value**; `86` in **Maximum value**.

  In order to identify repeated values, you may have found it helpful to tabulate the values in column `C1`. This is done by entering `C1` in the **Variables** field of the **Tally Individual Values** dialogue box (**Stat** > **Tables** > **Tally Individual Values**) and ensuring that **Count** is selected.

  When a member of the module team did this, the 30 random numbers given by Minitab were as follows.

```
77  30  84  86  17  65  32  30  75  58
80  39  11  75  86  24  68  51  65   4
10   9  58  44  67  72   8  65  56  83
```

  The tabulation of these numbers highlighted that in this set of numbers, 30, 58, 75 and 86 occur twice and 65 occurs three times. So the first 15 unique numbers are as follows.

```
77  30  84  86  17  65  32  75  58  80
39  11  24  68  51
```

The labels that the member of the module team obtained using Method 2 correspond to the following staff members.

Pat Trumpington	Gerald Flint	Paul Woodhouse
Tara Yeo	Jim Clarke	Andrew Pinder
Abraham Franks	Anna Thompson	Dick Masterton
Babs Tyndale	Maggie Greenway	Max Bramley
Emma Damper	Dan Ricardo	Chris Lang

Whichever method you used, you probably ended up with a different sample of staff members.

(b) Of the individuals in the sample given in the solution to part (a), 40% are female and 60% are male. This compares favourably with the target population (41% female and 59% male). The sample was unrepresentative in terms of occupation. In the sample, 53% are professional staff compared with 65% in the target population. Administrative staff, 13% of the sample, and secretarial staff, 27% of the sample, are slightly over-represented in the sample. (In the target population, 8% are administrative staff and 21% are secretarial staff.)

Solution to Computer activity 46

(a) The sum of the residuals increases when the line is moved downwards. As the line is moved up, the sum of the residuals decreases.

(b) Yes, when the sum of the residuals is zero, the line goes through the point $(\overline{x}, \overline{y})$.

(c) Whatever the slope of the line, the line goes through the point $(\overline{x}, \overline{y})$ when the sum of the residuals is zero.

Solution to Computer activity 47

(a) As the line becomes steeper, the sum of the squared residuals gets larger.

(b) As the line becomes less steep, the sum of the squared residuals initially reduces. Then the sum of the squared residuals increases again.

(c) When the sum of the squared residuals is minimised, the equation of the line is $y = 7 + 1.4x$. This line appears to provide a reasonable fit to the data.

(d) You should have found that the two lines have the same position, indicating that the line you found in part (c) is the least squares regression line.

Solution to Computer activity 48

(a) The output produced by Minitab includes the following:

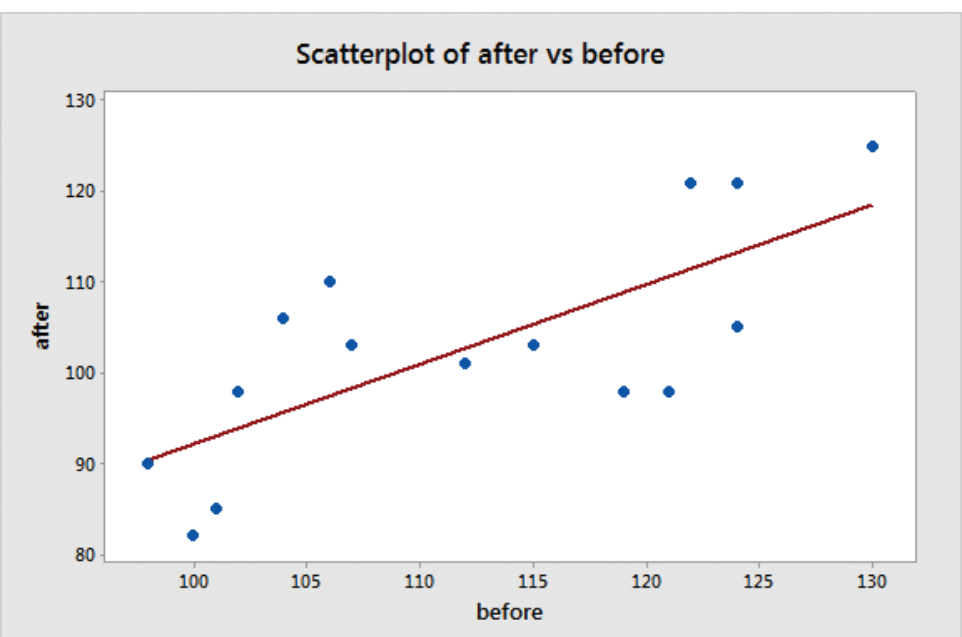
```
Regression Equation

after = 4.2 + 0.880 before
```

So the equation of the least squares regression line found by Minitab is $y = 4.2 + 0.880x$, where y is the blood pressure (in mmHg) two hours after injection and x is the blood pressure (in mmHg) before injection.

(b) The equation of the regression line given by Minitab is the same as the equation found at the end of Activity 17 in Unit 5.

Solution to Computer activity 49

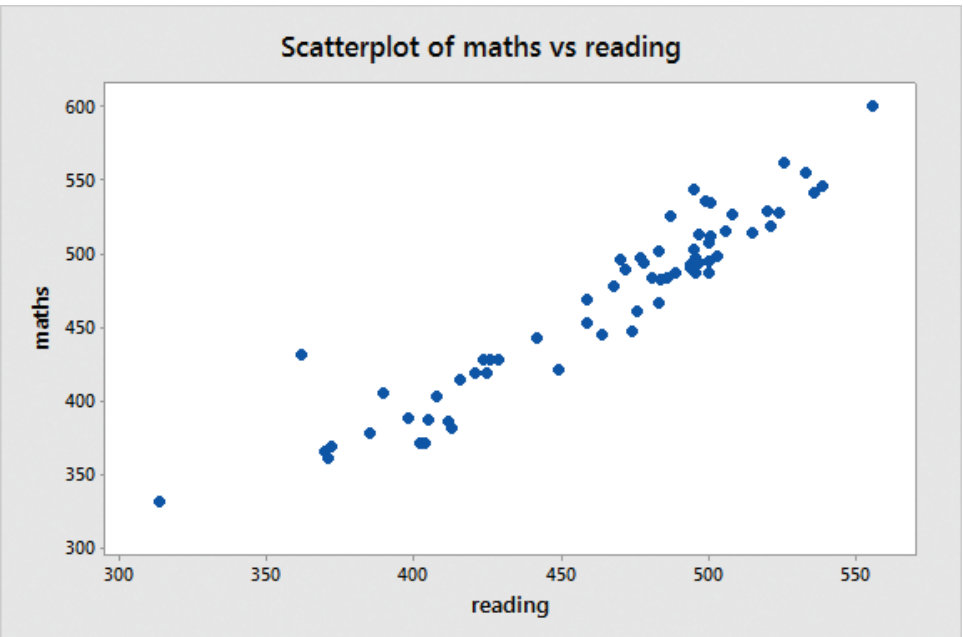
The scatterplot produced by Minitab with the least squares regression line on it is given below.



The line appears to be the same as the one given in the solution to Activity 17 of Unit 5 (Subsection 4.2). For example, the point with the lowest 'before' measure is just slightly below the line, and the line is almost parallel to two points in the middle of the data.

Solution to Computer activity 50

(a) The scatterplot produced by Minitab is shown below.



This was obtained by entering `maths` in the **Y variables** field and `reading` in the **X variables** field in the **Scatterplot: Simple** dialogue box (**Graph** > **Scatterplot** > **Simple**).

(b) The relationship between student achievement on the reading and mathematics scales is positive, linear and reasonably strong. One outlier is obvious: a country where the student achievement on the reading scale is low, but the achievement on the mathematics scale is high, given the achievement on the reading scale.

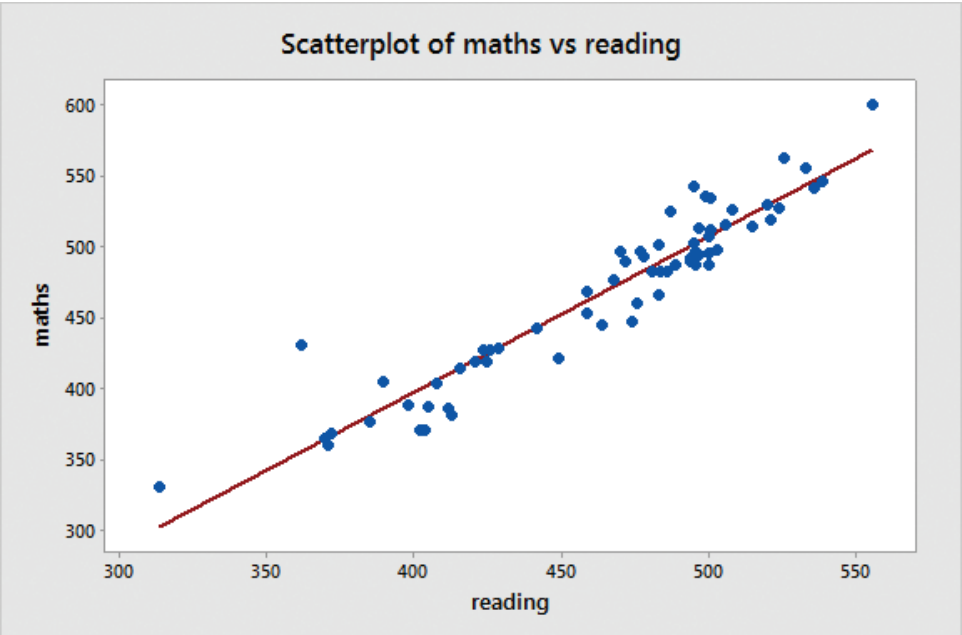
(c) The output produced by Minitab includes the following:

```
Regression Equation

maths = -42.7 + 1.0990 reading
```

This was obtained by entering `maths` in the **Responses** field and `reading` in the **Continuous predictors** field of the **Regression** dialogue box (**Stat** > **Regression** > **Regression** > **Fit Regression Model**).

So the equation of the least squares fit line is $y = -42.7 + 1.0090x$, where y is the score on the mathematics scale and x is the score on the reading scale.

(d) The scatterplot with the least squares fit line is given below.

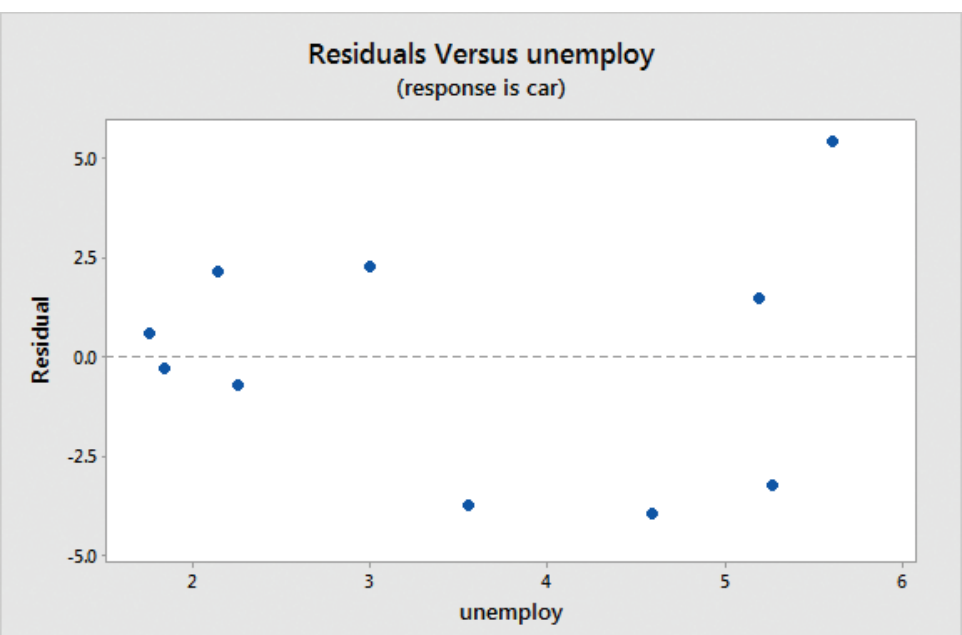


This was obtained using the **Add Regression Fit** dialogue box (**Editor** > **Add** > **Regression Fit**), making sure that **Linear** and **Fit intercept** were selected.

The line appears to fit the data reasonably well. It is generally in the middle of the pattern of points.

Solution to Computer activity 51

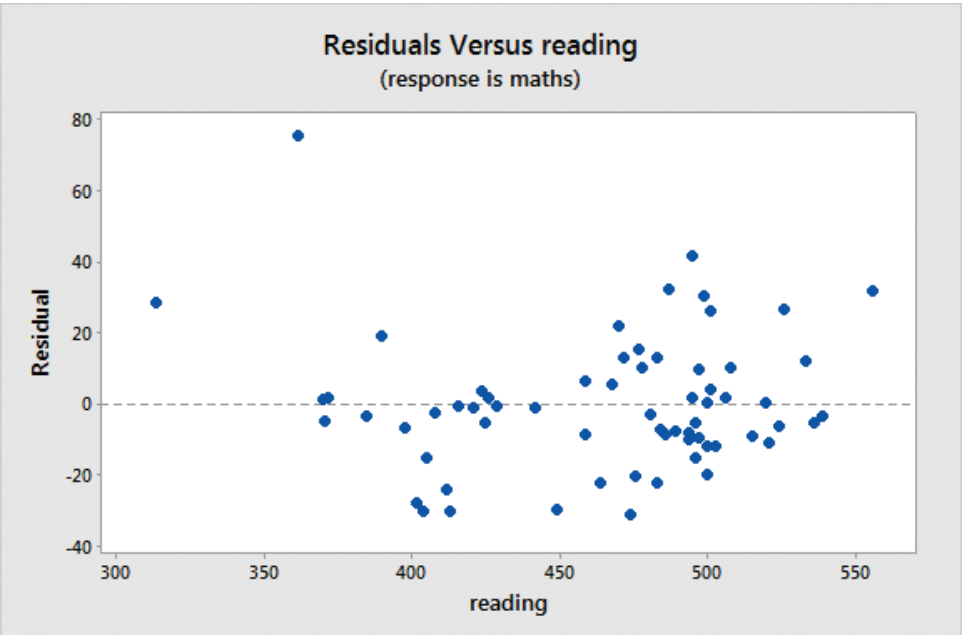
The residual plot produced by Minitab is given below.



The plot matches that given in Figure 49 of Unit 5 (Subsection 5.1). Only the scales used for the x- and y-axes differ.

Solution to Computer activity 52

(a) The residual plot produced by Minitab is given below.



This is obtained by entering `maths` in the **Responses** field of the **Regression** dialogue box (**Stat** > **Regression** > **Regression** > **Fit Regression Model**) and `reading` in the **Continuous predictors** field. Then in the **Regression: Graphs** dialogue box (obtained by clicking on **Graphs...** in the **Regression** dialogue box), select the `Regular` option and enter `reading` in the **Residuals versus the variables** field.

(b) The residual plot indicates that the fit of the least squares regression line is generally reasonable. Most of the points appear to be scattered around the line $y = 0$ with no obvious pattern. However, there is one residual that is much bigger than all the others (a residual of nearly 80 while all the others are roughly in the range minus 40 to plus 40). For this country – the country with the second worst student performance on the reading scale – student performance on the mathematics scale is much higher than its fitted value from the model.

Solution to Computer activity 55

(a) $P(x \leq 4) = 0.193\,85$ and $P(x \leq 5) = 0.387\,21$.

(b) The values go up as you go down this column. In other words, the cumulative probabilities increase as the value of x increases. This is to be expected as $P(x \leq 1) = P(x \leq 0) + P(x = 1)$, $P(x \leq 2) = P(x \leq 1) + P(x = 2)$, and so on, and the probabilities are all positive. (Probabilities can be zero or positive, but these probabilities are all positive.) So, positive values are added as we move down the list of cumulative probabilities.

(c) $P(x \leq 12) = 1$. Hence it is a certainty that 12 or fewer values will be above the median. This makes sense because there are 12 values and so having 12 or fewer values above the median covers all the possibilities for the number of values above the median.

Solution to Computer activity 56

(a) The cumulative probabilities x, with probabilities rounded to five decimal places, are as follows.

↓	C1	C2	C3	C4
	x	cumprob		
1	0	0.00001		
2	1	0.00014		
3	2	0.00117		
4	3	0.00636		
5	4	0.02452		
6	5	0.07173		
7	6	0.16615		
8	7	0.31453		
9	8	0.50000		
10	9	0.68547		
11	10	0.83385		
12	11	0.92827		
13	12	0.97548		
14	13	0.99364		
15	14	0.99883		
16	15	0.99986		
17	16	0.99999		
18	17	1.00000		
19				

This was obtained by first creating a worksheet with a column called `x` containing the numbers from 0 to 17 (**File** > **New** and then **Calc** > **Make Patterned Data** > **Simple Set of Numbers**).

Then, in the **Binomial Distribution** dialogue box (**Calc** > **Probability Distributions** > **Binomial**), the **Cumulative probability** option should be selected. Also, `17` should be entered in the **Number of trials** field, `0.5` in the **Event probability** field, `x` in the **Input column** field and `cumprob` in the **Optional storage** field.

(b) $P(x \leq 4)$ is the cumulative probability when $x = 4$ and $P(x \leq 5)$ is the cumulative probability when $x = 5$. So $P(x \leq 4) = 0.024\,52$ and $P(x \leq 5) = 0.071\,73$.

(c) Suppose we wanted to test the hypothesis that the population median usable life is 60 months. If we observed that in a sample of 17 batteries only 4 batteries had a usable life above 60 months, then the p-value from the sign test would be

$$\begin{aligned} &P(4 \text{ or fewer } [+] \text{ values }) + P(4 \text{ or fewer } [-] \text{ values}) \\ &\quad = 2 \times P(x \leq 4) \\ &\quad = 2 \times 0.024\,52 \quad \text{(from (b))} \\ &\quad = 0.049\,04. \end{aligned}$$

This p-value is less than 0.05 so the hypothesis would be rejected at the 5% significance level, and hence 4 is in the critical region. On the other hand,

$$\begin{aligned} &P(5 \text{ or fewer } [+] \text{ values}) + P(5 \text{ or fewer } [-] \text{ values}) \\ &\quad = 2 \times P(x \leq 5) \\ &\quad = 2 \times 0.071\,73 \\ &\quad = 0.143\,46, \end{aligned}$$

which is greater than 0.05.

Thus the hypothesis that the median usable life is 60 months would not be rejected at the 5% significance level if 5 of the 17 batteries had a usable life above 60 months.

In summary, the hypothesis is rejected at the 5% significance level if $x = 4$ but not if $x = 5$. Hence 4 is the critical value, in agreement with Table 8 of Unit 6.

Solution to Computer activity 58

(a) After having made sure that **petrol3.mtw** is the active worksheet in Minitab, obtain the **1-Sample Sign** dialogue box (**Stat** > **Nonparametrics** > **1-Sample Sign**). In the **1-Sample Sign** dialogue box, enter `mpg` in the **Variables** field and `36` in the **Test median** field, and make sure that `not equal` is given in the **Alternative** field.

The resulting output produced by Minitab is as follows.

```
Sign test of median =  36.00 versus ≠ 36.00

      N  Below  Equal  Above       P  Median
mpg  34     20      0     14  0.3915   35.34
```

So the p-value from the test is 0.3915. This is large – much bigger than 0.10. So from Table 1 there is little evidence against the hypothesis. That is, there is little evidence against the claim that the car's petrol consumption is 36 miles per gallon.

(b) Obtain the **1-Sample Sign** dialogue box again. This time change the **Test median** field to 37. The output provided by Minitab this time is as follows.

```
Sign test of median =  37.00 versus ≠ 37.00

       N  Below  Equal  Above       P  Median
mpg   34     26      0      8  0.0029   35.34
```

From the Minitab output, the p-value from the test is now 0.0029. This is less than 0.01, but greater than 0.001. So using Table 1, there is strong evidence against the hypothesis. That is, there is strong evidence that the car's median petrol consumption is not 37 miles per gallon. In fact, it appears to be less than 37 miles per gallon.

Solution to Computer activity 59

(a) When you change the value from $\mu = 0$ to $\mu = 1$, the normal distribution has the same shape but moves to the right, giving the same output as Figure 31(b).

(b) When you change the value from $\mu = 0$ or $\mu = 1$ to $\mu = -1$, the normal distribution has the same shape but moves to the left, giving the same output as Figure 31(e).

(c) As you vary the value for μ over the range allowed, the shape of the normal distribution remains the same. However, the normal distribution moves to the right of its starting point ($\mu = 0$) if $\mu > 0$. The larger the value of μ, the further it moves. On the other hand, the normal distribution moves to the left of its starting point ($\mu = 0$) if $\mu < 0$. This time the smaller (more negative) the value of μ, the further it moves.

Solution to Computer activity 60

(a) When you change the value from $\sigma = 1$ to $\sigma = 2$, the normal distribution stays at the same location but it becomes more spread out, giving the same output as Figure 32(c).

(b) When you change the value from $\sigma = 1$ or $\sigma = 2$ to $\sigma = 0.5$, the normal distribution still stays at the same location but it becomes less spread out, and more peaked, giving the same output as Figure 32(a).

(c) As you vary the value of σ over the range allowed, the location of the normal distribution remains the same. However, the normal distribution becomes more spread out compared with its starting point ($\sigma = 1$) if $\sigma > 1$, becoming flatter and flatter as the value of σ increases. On the other hand, the normal distribution becomes less spread out compared with its starting point ($\sigma = 1$) if $\sigma < 1$, becoming more and more peaked as the value of σ decreases.

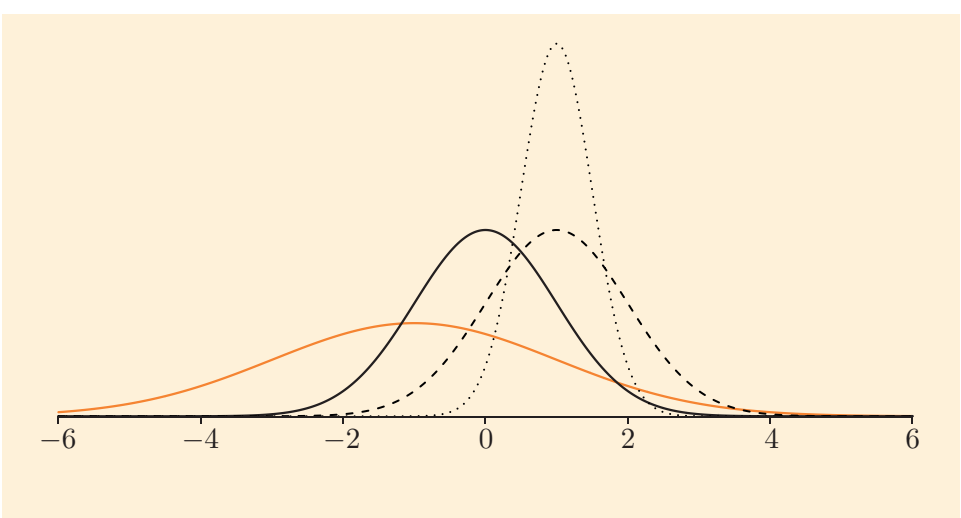
Solution to Computer activity 61

(a) When you change the mean from $\mu = 0$ to $\mu = 1$, the normal distribution moves to the right. When you change the standard deviation from $\sigma = 1$ to $\sigma = 2$, the normal distribution also becomes more spread out. So the normal distribution corresponding to $\mu = 1$ and $\sigma = 2$ is both moved to the right and flattened out, relative to the normal distribution with $\mu = 0$ and $\sigma = 1$.

(b) Statement A is incorrect. The incorrect claim is that the normal distribution moves to the left, when actually the normal distribution moves to the *right* when $\mu > 0$. A corrected version of statement A is therefore: 'When $\mu > 0$ and $\sigma < 1$, the normal distribution is moved to the right and is less spread out relative to the normal distribution with $\mu = 0$, $\sigma = 1$.' An example is the normal distribution with $\mu = 1$ and $\sigma = 0.5$ shown (as a dotted line) in the figure below.

Statement B is correct. An example is the normal distribution with $\mu = 1$ and $\sigma = 1$ shown (as a dashed line) in the figure below.

Statement C is incorrect. The incorrect claim is that the normal distribution is less spread out, when actually the normal distribution is more spread out when $\sigma > 1$. A corrected version of statement C is therefore: 'When $\mu < 0$ and $\sigma > 1$, the normal distribution is moved to the left and is more spread out relative to the normal distribution with $\mu = 0$, $\sigma = 1$.' An example is the normal distribution with $\mu = -1$ and $\sigma = 2$ shown (on the left) in the figure below.



Normal distributions with $\mu = -1$, $\sigma = 2$, to the left, $\mu = 0$, $\sigma = 1$, in the centre, and $\mu = 1$, $\sigma = 0.5$ (dotted) and $\mu = 1$, $\sigma = 1$ (dashed), to the right

Solution to Computer activity 62

(a) Setting $a = 1$ shifts this distribution to the left so that its mode occurs at zero.

(b) Setting $b = 2$ rescales this distribution to match the standard normal distribution.

Solution to Computer activity 63

(a) For this normal distribution, setting $a = 2.5$ shifts it to the left so that its mode occurs at 0, and setting $b = 0.5$ rescales the distribution to match the standard normal distribution.

(b) For this normal distribution, setting $a = -1.5$ shifts it to the right so that its mode occurs at zero, and setting $b = 1.5$ rescales the distribution to match the standard normal distribution.

(c) In each part of this activity, the values of a and b needed to transform a normal distribution with mean μ and standard deviation σ to the standard normal distribution are $a = \mu$ and $b = \sigma$. So the formula to transform a normal distribution with mean μ and standard deviation σ to a standard normal distribution does appear to be $z = (x - \mu)/\sigma$.

Solution to Computer activity 65

The null and alternative hypotheses are

H_0: $\mu = 598$

H_1: $\mu \neq 598$,

where μ is the population mean weekly wage (in £) of male leisure and sports managers in 2011.

As the sample size, 230, is greater than 25, it is appropriate to apply the z-test.

In Minitab, do the following.

- Click on **Stat**, choose **Basic Statistics**, and then choose **1-Sample Z...**. The **One-Sample Z for the Mean** dialogue box appears.
- Make sure the `Summarized data` option is selected in the top drop-down list. Type `230` in the **Sample size** field and `587` in the **Mean** field.
- Type `208` in the **Known standard deviation** field.
- Make sure that **Perform hypothesis test** is selected. Type `598` in the **Hypothesized mean** field.
- Click on **OK**.

The resulting Minitab output is shown below.

```
Test of μ = 598 vs ≠ 598
The assumed standard deviation = 208

   N   Mean  SE Mean       95% CI            Z       P
230  587.0     13.7  (560.1, 613.9)   -0.80  0.423
```

The p-value is 0.423 which, according to Table 1 (Subsection 6.2), gives little evidence against the null hypothesis. Equivalently, the null hypothesis is not rejected at the 5% level. (These claims follow because $p = 0.423 > 0.05$.) So there is little evidence that the mean weekly wage of

male leisure and sports managers differed from the overall mean weekly wage for male employees in 2011.

Solution to Computer activity 66

The null and alternative hypotheses are

H_0: $\mu = 255$

H_1: $\mu \neq 255$,

where μ is the population mean weight of glass per milk bottle.

As the sample size, 27, is greater than 25, it is appropriate to apply the z-test.

In Minitab, do the following.

- Obtain the **One-Sample Z for the Mean** dialogue box (**Stat** > **Basic Statistics** > **1-Sample Z**).
- Make sure the `Summarized data` option is selected in the top drop-down list. Type `27` in the **Sample size** field and `256.19` in the **Mean** field.
- Type `2.5` in the **Known standard deviation** field.
- Make sure that **Perform hypothesis test** is selected. Type `255` in the **Hypothesized mean** field.
- Click on **OK**.

The resulting Minitab output is shown below.

```
Test of μ = 255 vs ≠ 255
The assumed standard deviation = 2.5


 N     Mean   SE Mean         95% CI            Z       P
27  256.190     0.481  (255.247, 257.133)   2.47   0.013
```

The p-value is 0.013 which, according to Table 1 (Subsection 6.2), gives moderate evidence against the null hypothesis. Equivalently, the null hypothesis is rejected at the 5% level but not rejected at the 1% level. (These claims follow because $0.05 \geq p = 0.013 > 0.01$.) We conclude that there is some evidence that the mean weight of the glass per milk bottle has changed and perhaps the machine should be adjusted.

Solution to Computer activity 67

The six numbers in Table 3 are displayed in the worksheet in three rows and two columns. The labels for the rows ('Analytic phonics', 'Analytic phonics + PA' and 'Synthetic phonics') are also given in the main body of the worksheet – in the first column. However, the labelling of the columns is *not* shown in the main body of the worksheet. The column headings 'Not higher' and 'Higher' are used as variable names, and the label 'Reading age compared to chronological age' does not appear at all.

Note that the first column is labelled `C1-T`; the `T` in `C1-T` stands for Text, indicating that the entries are text, not numbers.

The marginal totals have not been entered in the worksheet. This is how it should be. Entering the marginal totals into the worksheet could lead to errors in the χ^2 test, as Minitab has no way of distinguishing these totals from the other counts. And in any case, as you will see in Computer activity 68, Minitab will calculate the marginal totals for you.

Solution to Computer activity 69

(a) The Expected value for the `higher` category in the `Analytic phonics` group is 49.15 and the χ^2 contribution from the `nothigher` category in the `Synthetic phonics` group is 10.151. The χ^2 test statistic is 35.158.

These were read off from the following output in the Session window.

```
Rows: Teaching group   Columns: Worksheet columns

                        nothigher  higher  All

Analytic phonics               68      36  104
                            54.85   49.15

Analytic phonics + PA          51      24   75
                            39.55   35.45

Synthetic phonics              35      78  113
                            59.60   53.40

All                           154     138  292

Cell Contents:      Count
                    Expected count


Pearson Chi-Square = 35.158, DF = 2, P-Value = 0.000
Likelihood Ratio Chi-Square = 35.855, DF = 2, P-Value = 0.000
```

(Note that the Expected values match those given in Table 23 of Unit 8 (Subsection 4.2).)

The output was obtained by first opening the **firstR.mtw** worksheet. Then, in the **Chi-Square Test for Association** dialogue box (**Stat** > **Tables** > **Chi-Square Test for Association**), making sure `Summarized data in a two-way table` was selected in the top drop-down list, entering `nothigher` and `higher` into the **Columns containing the table** field, and entering `'Teaching group'` in the **Row** field, before clicking on **OK**.

(b) The p-value is reported as 0.000, which means $p < 0.0005$. (Remember, the p-value is rounded to three decimal places so it does *not* mean $p = 0$.) From Table 1 (Subsection 6.2), there is very strong evidence against the null hypothesis. Also, since $p \leq 0.01$, the null hypothesis is rejected at the 1% significance level (and hence also is rejected at the 5% significance level). Thus, we conclude that there is very strong evidence of an association between reading ability at the first follow-up test and the teaching method for children starting primary school.

Solution to Computer activity 70

(a) The output from the χ^2 test is as follows.

```
Rows: Teaching group   Columns: Worksheet columns

                         lowlow  lowhigh  highlow  highhigh  All

Analytic phonics             50       16       17        20  103
                          36.46    27.61    17.70     21.24

Analytic phonics + PA        30       10       21        14   75
                          26.55    20.10    12.89     15.46

Synthetic phonics            23       52       12        26  113
                          40.00    30.29    19.42     23.30

All                         103       78       50        60  291

Cell Contents:      Count
                    Expected count


Pearson Chi-Square = 46.716, DF = 6, P-Value = 0.000
Likelihood Ratio Chi-Square = 46.374, DF = 6, P-Value = 0.000
```

This was obtained by making sure `Summarized data in a two-way table` was selected from the top drop-down list in the **Chi-Square Test for Association** dialogue box (**Stat** > **Tables** > **Chi-Square Test for Association**), entering `lowlow`, `lowhigh`, `highlow` and `highhigh` into the **Columns containing the table** field and entering `'Teaching group'` in the **Rows** field.

The degrees of freedom given by Minitab are 6. The formula gives $(3 - 1) \times (4 - 1) = 6$, since the table comprises 3 rows and 4 columns. So these two values match.

(b) The χ^2 test statistic is 46.716 and the p-value is reported as 0.000, which means $p < 0.0005$. Thus, from Table 1 (Subsection 6.2), there is very strong evidence against the null hypothesis. Also, since $p \leq 0.01$, the null hypothesis is rejected at the 1% significance level. Now, the null hypothesis for this test is that the teaching method for children starting primary school and the reading abilities at baseline/first follow-up are independent. Thus, we conclude that there is very strong evidence of an association between reading abilities in the baseline/first follow-up tests and teaching method for children starting primary school.

Solution to Computer activity 71

The three cells with Expected values less than 5 are in the `lowhigh` column. One way round the problem is to combine the `lowhigh` and `highlow` categories into a single category, which would represent a sort of 'mixed ability' group, intermediate between `lowlow` and `highhigh`. An alternative is to combine the `lowlow` and `lowhigh` columns, though this then loses the 'mixed ability' interpretation of the new combined category.

Solution to Computer activity 72

(a) The new category `mixed` has counts 14, 6 and 9. The Minitab worksheet should look as below. (This worksheet has also been saved as **secondSR2.mtw**.)

↓	C1-T	C2	C3	C4	C5
	Teaching group	lowlow	mixed	highhigh	
1	Analytic phonics	7	14	74	
2	Analytic phonics + PA	6	6	54	
3	Synthetic phonics	8	9	88	
4					

(b) The χ^2 test is performed as described in Computer activity 68. There is now no warning message in the Session window; the lowest Expected value is 5.21. Thus, combining the columns has successfully dealt with the problem of low Expected values.

The penultimate line of the output in the Session window should read

```
Pearson Chi-Square = 2.385, DF = 4, P-Value = 0.665
```

Thus, there is little evidence against the null hypothesis, since $p > 0.10$, and it is not rejected at the 5% level, since $p > 0.05$. In other words, children's spelling and reading abilities at the second follow-up tests appear to be unrelated to which teaching method they were originally allocated.

Solution to Computer activity 74

(a) The correlation coefficient between `english` and `maths` is 0.702. This was obtained by adding `english` and `maths` to the **Variables** field in the **Correlation** dialogue box (**Stat** > **Basic Statistics** > **Correlation**).

The correlation coefficient is positive and reasonably close to 1, which means there is a reasonably strong positive linear relationship between the pass rate in English and the pass rate in Mathematics.

(b) The correlation coefficients are as follows:

- English and Science: 0.429
- English and History or Geography: 0.674
- English and Languages: 0.281.

The relationship between the pass rates for English and Mathematics is the strongest as the corresponding correlation coefficient is the closest to +1 or −1. The relationship between the pass rates for English and Languages is the weakest as the corresponding correlation coefficient is the closest to 0.

(c) The correlation between `english` and itself is 1.000, which is not surprising. There must be a perfect linear relationship between a variable and itself, so the correlation coefficient necessarily takes the value +1.

Solution to Computer activity 76

(a) Using Minitab gives the mean as 458 g and the standard deviation as 6.18 g.

This information was obtained by entering `weight` in the **Variables** field of the **Display Descriptive Statistics** dialogue box (**Stat** > **Basic Statistics** > **Display Descriptive Statistics**).

(b) The 95% confidence interval for the mean weight of jam is (456.01 g, 459.99 g). This is slightly different to the interval you calculated in Activity 18 of Unit 9 (Subsection 4.2) because the standard deviation was only given to two decimal places.

The interval was obtained by selecting `Summarized data` from the drop-down list in the **One-Sample Z for the Mean** dialogue box (**Stat** > **Basic Statistics** > **1-Sample Z**), then entering `37` in the **Sample size** field, `458` in the **Mean** field and `6.18` in the **Standard deviation** field. You also needed to ensure that `95` was entered in the **Confidence level** field of the **One-Sample Z: Options** dialogue box.

(c) The 99% confidence interval is (455.38 g, 460.62 g).

This interval was obtained in the same way as the interval in part (b). However, this time the value `99` needed to be entered in the **Confidence level** field of the **One-Sample Z: Options** dialogue box.

(d) The weight 454 g is not in either of the confidence intervals, so there is strong evidence that the mean weight of jam in the jars is not 454 g. However, as both confidence intervals are actually above 454 g, it is unlikely that any purchasers of the jam will complain – on average they are getting a little bit more jam than they are led to believe they will!

Solution to Computer activity 77

(a) It is likely that both the slope and intercept of your least squares regression line are different to the population line.

(b) Whether or not your second line is closer to the population depends on the closeness of your first line (in part (a)). The further away your first line was, the more likely it is that your second line is closer to the population line.

(c) The least squares regression lines make up a band on the plot. This band is narrowest in the middle and widest at the two ends. The population line appears to be in the middle of the band. So, on average, the least squares regression lines are similar to the population line. However, all we can say about any one line is that generally it lies within a band on the plot.

Solution to Computer activity 78

(a) The answer depends on your data. However, the centre of your interval is likely to have been around 5 and the width is likely to have been about 0.6.

(b) The population line is $y = 2.5 + 0.5x$. So when $x = 5$, the Expected value of y is $2.5 + 0.5 \times 5 = 5$. Therefore, it is likely that your interval contained the value $y = 5$.

(c) The confidence intervals are likely to be different. It is likely that the endpoints are not the same, and the widths differ and so do the centres of the intervals. However, this interval probably also contained the value y_{true}.

(d) Overall you should have found that around 95 of the intervals contained the value y_{true}. This is to be expected as each interval corresponds to a 95% confidence interval for the mean response. In other words, these intervals are calculated in such a way that for any particular value of x, 95% of them on average will contain the corresponding value of y_{true}.

Solution to Computer activity 79

(a) You should notice that the variability along the x-axis does not alter very much. However, as the strength of the relationship reduces, the points get more and more scattered around the line.

(b) When the relationship is weak, you should find that the confidence interval is much wider than the interval when the relationship is moderate. When the relationship is strong, you should find that the interval is much narrower than when the relationship is moderate.

(c) You should find in each case that about 95 of the intervals contain y_{true}. So although the intervals are different widths, they are each just as likely to contain y_{true}.

Solution to Computer activity 80

(a) You should find that the intervals tend to be about 0.6 units wide. However, as you found in Computer activity 78, the width does vary a bit from sample to sample.

(b) You should have found that about 95% of the intervals contain the correct value.

(c) Changing the slope of the line from 0.3 to 10 changes the value around which each interval is centred. However, the width of the interval, and the proportion of intervals which contain the correct value, should remain about the same.

(d) Similarly to part (c), changing the slope of the line to -5 only changes the value around which each interval is centred. The width of the interval, and the proportion of intervals which contain the correct value, should remain about the same.

(e) You should have found that the slope of the line only makes a difference to the value about which a confidence interval is centred. The width of the interval is not affected by the value of the slope.

Solution to Computer activity 81

(a) You should find that the confidence intervals are roughly 0.6 units wide and that about 95% of them contain the correct value.

(b) When the sample consists of 40 points, the confidence intervals corresponding to $x = 5$ are roughly 0.3 units wide. So increasing the sample size has more than halved the width of the confidence interval. However, again, about 95% of them contain the correct value.

(c) When the sample consists of 160 points, the confidence intervals corresponding to $x = 5$ are roughly 0.15 units wide. Again, increasing the sample size has more than halved the width of the confidence interval. However, again, about 95% of them contain the correct value.

Solution to Computer activity 83

(a) The equation of the regression line given by Minitab is `KS4 = -36.4 + 1.098 KS2`. This was obtained by entering `KS4` in the **Responses** field of the **Regression** dialogue box (**Stat** > **Regression** > **Regression** > **Fit Regression Model**) and entering `KS2` in the **Continuous predictors** field.

(b) The 95% confidence interval for the mean value of `KS4` for constituencies like Blaenau Gwent is 31.7% to 44.8% to one decimal place.

This was obtained by entering `68.0` in the **KS2** field of the **Predict** dialogue box (**Stat** > **Regression** > **Regression** > **Predict**).

The confidence interval means that we can be 95% confident that the statement 'the population mean percentage of Key Stage 4 students attaining this benchmark in 2011 in constituencies like Blaenau Gwent is between 31.7% and 44.8%' is true.

(c) The 95% prediction interval for the value of `KS4` for a constituency like Caerphilly is 28.4% to 58.1% to one decimal place.

This was obtained using the same procedure as in part (b), except this time entering `72.5` in the **KS2** field of the **Predict** dialogue box.

The prediction interval means that we would expect the statement 'the percentage of Key Stage 4 students attaining this benchmark in 2011 in a constituency like Caerphilly is between 28.4% and 58.1%' to be true 95% of the time.

(d) The prediction intervals are as follows.

Observed values and prediction intervals for the values of `KS4`

Constituency	`KS4`	95% prediction interval for `KS4`
Blaenau Gwent	35.3%	22.8% to 53.7%
Caerphilly	37.0%	28.4% to 58.1%
Islwyn	50.7%	26.1% to 56.0%
Merthyr Tydfil and Rhymney	38.2%	22.5% to 53.6%
Monmouth	53.4%	35.7% to 69.2%
Newport East	36.4%	27.9% to 57.6%
Newport West	54.5%	36.1% to 70.1%
Torfaen	41.2%	22.3% to 53.4%

The prediction intervals, as a whole, do look reasonable. All of the actual values for `KS4` are inside their respective 95% prediction intervals. With a small sample, this is to be expected. Only 5% of values would be expected to be outside the corresponding interval – just 0.4 values in this case. However, with larger samples, it is not good for all of the actual values to lie within their respective 95% prediction intervals, as this would indicate that the intervals are too wide.

Solution to Computer activity 85

(a) The confidence interval is obtained by selecting `One or more samples, each in a column` in the top drop-down list and entering `speed` in the next field of the **One-Sample t for the Mean** dialogue box. Note that the **Perform hypothesis test** option does not need to be selected.

The 95% confidence interval is given as `(709.9, 802.5)` and hence the 95% confidence interval for the speed of light in air (in km/s) is (299 709.9, 299 802.5).

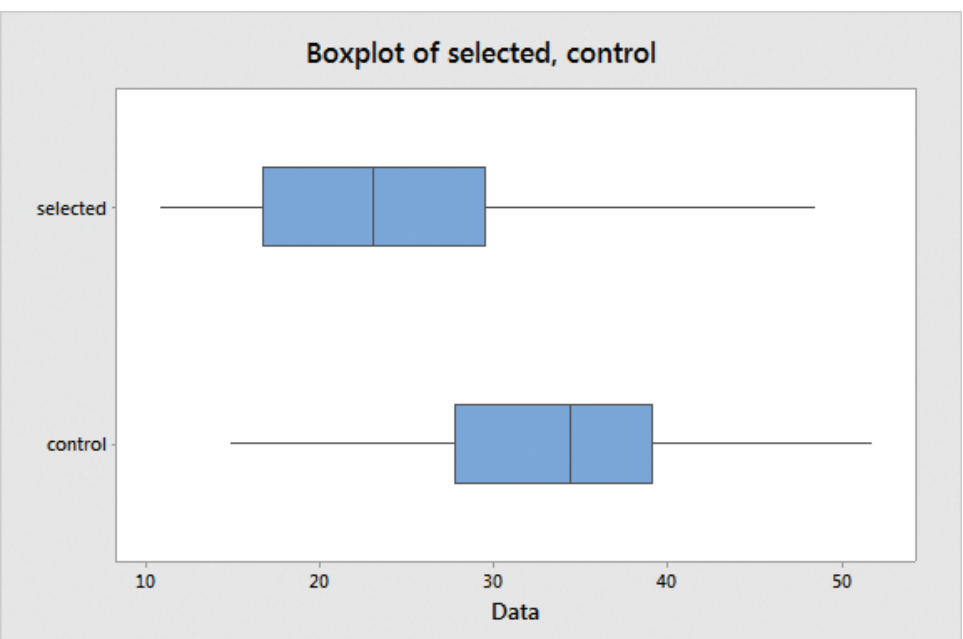
(b) This time in the **One-Sample t for the Mean** dialogue box, the **Perform hypothesis test** option needs to be selected and `705` entered in the **Hypothesized mean** field.

The value of the test statistic (`T` in the output) is 2.29 and the p-value is 0.032. So, from Table 1 (Subsection 6.2), there is moderate evidence against the null hypothesis. This means there is moderate evidence that the claim $\mu = 705$ is not correct – at least for the conditions which applied when Michelson was taking his measurements. An alternative explanation for this result is that the method used by Michelson to measure the speed of light was slightly biased.

(c) The value 705 is not in the 95% confidence interval that was given by Minitab. So the p-value for the one-sample t-test must be less than 0.05 – as found in part (b).

Solution to Computer activity 87

(a) The diagram containing both boxplots is as follows.



These boxplots were obtained by entering both `selected` and `control` in the **Graph variables** field in the **Boxplot: Multiple Y's, Simple** dialogue box (**Graph** > **Boxplot** and selecting **Multiple Y's, Simple**). Also, make sure that the **Transpose value and category scales** option is selected in the **Boxplot: Scale** dialogue box (obtained by clicking on the **Scale...** button).

Both boxplots appear to be reasonably symmetric and there are no potential outliers indicated. So it is reasonable to assume that the data in both groups are samples from normal population distributions.

(b) The output produced by Minitab is as follows.

```
Two-sample T for selected vs control

           N   Mean  StDev  SE Mean
selected  50  24.44   8.78      1.2
control   25  33.37   8.94      1.8


Difference = μ (selected) - μ (control)
Estimate for difference:  -8.93
95% CI for difference:  (-13.24, -4.62)
T-Test of difference = 0 (vs ≠): T-Value = -4.13  P-Value = 0.000  DF = 73
Both use Pooled StDev = 8.8303
```

This was obtained by selecting `Each sample is in its own column` from the top drop-down list of the **Two-Sample t for the Mean** dialogue box (**Stat** > **Basic Statistics** > **2-Sample t**) and entering `selected` and `control` in the **Sample 1** and **Sample 2** fields respectively. Also by making sure that **Assume equal variances** was selected in the **Two-Sample t: Options** dialogue box.

So the test statistic is -4.13 and the corresponding p-value is given as 0.000. (Remember that a p-value of 0.000 means that $p < 0.0005$, not that p is exactly equal to zero.) This means that from Table 1 (Subsection 6.2), there is very strong evidence against the null hypothesis. In other words, there is very strong evidence that the egg-laying capabilities in the two groups of fruit flies are not the same. In fact, the egg-laying capability of selectively bred fruit flies appears to be lower than that of fruit flies that have not been selectively bred.

(c) The standard deviation of eggs laid in the selective breeding group is given as 8.78, and hence the variance is $8.78^2 = 77.0884$. Similarly, the variance of eggs laid in the group that had not been selectively bred is 79.9236. The ratio of the bigger variance to the smaller is much less than 3. So, by the rule of thumb introduced in Subsection 3.3 of Unit 10, it is reasonable to have used the pooled standard deviation for the test completed in part (b).

(d) The 95% confidence interval for $\mu_s - \mu_c$ can be read off directly from the output obtained in part (b). It is $(-13.24, -4.62)$.

Solution to Computer activity 89

(a) The data consist of pairs of observations, one pair for each sample. The matched-pairs t-test is a good choice for such data because interest focuses on the difference within pairs, and not between pairs.

(b) The output from the matched-pairs t-test is as follows.

```
Paired T for A - B

               N     Mean   StDev  SE Mean
A             10    98.10   25.41     8.03
B             10   105.00   24.49     7.75
Difference    10    -6.90   10.87     3.44


95% CI for mean difference: (-14.67, 0.87)
T-Test of mean difference = 0 (vs ≠ 0): T-Value = -2.01  P-Value = 0.076
```

This output was obtained by selecting the `Each sample is in a column` from the drop-down list in the **Paired t for the Mean** dialogue box (**Stat** > **Basic Statistics** > **Paired t**) and entering `A` and `B` in the **Sample 1** and **Sample 2** fields, respectively. (You may have to remove `formL` and `formR` from the **Sample 1** and **Sample 2** fields first.)

The test statistic is therefore -2.01 and the corresponding p-value is 0.076. From Table 1 (Subsection 6.2) this means there is weak evidence that the two devices are not giving the same value on average.

(c) From the output obtained in part (b), the 95% confidence interval for the difference in readings is $(-14.67, 0.87)$. Note that this interval contains 0, which is the value that corresponds to no difference between the devices. This ties in with the conclusion you drew in part (b) – there is only weak evidence of a difference between the devices.

Solution to Computer activity 91

(a) The null and alternative hypotheses are

$$H_0\colon\ \mu = 6.8$$
$$H_1\colon\ \mu < 6.8,$$

where μ is the population mean sleeping time (in hours) of American fibromyalgia sufferers.

(b) The hypotheses just involve one group of patients, American fibromyalgia sufferers. So an appropriate test is a one-sample test – either the one-sample z-test or the one-sample t-test.

As the sample size, 744, is much greater than 25, it is appropriate to apply the z-test.

(c) To perform the test, do the following.

- In the **One-Sample Z for the Mean** dialogue box (**Stat** > **Basic Statistics** > **1-Sample Z**) select `Summarized data` from the top drop-down list.
- Type `744` in the **Sample size** field and `5.6` in the **Sample mean** field. Type `1.6` in the **Known standard deviation** field. Make sure that **Perform hypothesis test** is selected. Type `6.8` in the **Hypothesized mean** field.
- Click on **Options...** in the **One-Sample Z for the Mean** dialogue box.
- The alternative hypothesis corresponds to H_1: $\mu < 6.8$, so in the **One-Sample Z: Options** dialogue box, select `Mean < hypothesized mean` in the **Alternative hypothesis** field.
- Click on **OK** and on **OK** again.

The resulting Minitab output is shown below.

```
Test of μ = 6.8 vs < 6.8
The assumed standard deviation = 1.6


  N    Mean  SE Mean  99% Upper Bound       Z      P
744  5.6000   0.0587           5.7365  -20.46  0.000
```

The p-value associated with this one-sided z-test is, to three decimal places, 0.000. According to Table 1 (Subsection 6.2), this p-value, which is such that $0.001 \geq p$, gives very strong evidence against the null hypothesis. We conclude that there is very strong evidence that American fibromyalgia patients do indeed sleep for fewer hours per night, on average, than is the 'norm' for the entire US population.

Solution to Computer activity 92

(a) When a member of the module team did this, the following numbers were generated.

```
45  67  38  80   6  50  78  72  75  15
61  62  65  50  77   4  28  27  93  59
98  47  47  74  27  66  49  45  71  85
```

This was done by entering the following in the **Integer Distribution** dialogue box (**Calc** > **Random Data** > **Integer**): `30` in the **Number of rows of data to generate** field, `C1` in the **Store in column(s)** field, `1` in the **Minimum value** field and `100` in the **Maximum value** field.

The numbers you get are likely to be different to these.

(b) When a member of the module team did this, the following numbers were generated.

```
 6  44  59  58 100  77  54  55  94  72
42  92  94  92  30  57  21  42  11  70
 9  12  92  40  67  50  75  78  34  37
```

This time, because the random number generator seed was first set to take the value 10, you should have generated exactly the same numbers.

(c) You should have ended up with exactly the same numbers in column `C3` as in column `C2`. Repeating part (b), including starting with the same seed, results in Minitab generating exactly the same 'random' numbers.

(d) You should have ended up with the following numbers in column `C4`.

```
87  44  94  73  22   1  70  57  50  75
46  58  86  73   1  29  12  65  99  31
95  54  16  78  35  64  21  86  96  86
```

These numbers are different to those generated in parts (b) and (c) because Minitab has started in a different place in its sequence of pseudo-random numbers. However, it turns out that the position chosen to start at is decided in an entirely deterministic way. So your set of 'random' numbers should be exactly the same as those obtained by the module team (and anyone else who does this activity!).

Solution to Computer activity 93

(a) Overall, 50 of the participants were assigned to the control group and 50 were assigned to the experimental group.

(b) Yes there is balance. There are 25 individuals in each of the control and experimental groups within each centre.

(c) There is no randomness in the assignment of participants to treatments. So not setting the seed for the random number generator will not make any difference.

Solution to Computer activity 94

There are 55 participants assigned to the control group and 45 to the experimental group. In centre 1, there are 29 participants assigned to the control group and 21 to the experimental group. In centre 2, there are 26 participants assigned to the control group and 24 to the experimental group. The allocation of participants to groups in both centres is not balanced; in each centre the number of participants allocated to the experimental group does not exactly match the number of participants allocated to the control group. The number in the experimental group is particularly low in centre 1, whereas in centre 2 the allocation is roughly balanced as the numbers in the two groups are only a little bit different.

Solution to Computer activity 95

There are 50 participants in each of the control and experimental groups, so this is balanced. In centre 1, there are 22 participants assigned to the experimental group, and in centre 2, there are 28 participants assigned to the experimental group. So this method of randomisation does not produce exact balance within centres. The degree of balance is between those found for the two centres in Computer activity 94.

Solution to Computer activity 96

(a) The χ^2 test is appropriate because the data are categorical.

(b) The output produced by Minitab for this χ^2 test is as follows.

```
Rows: treatment   Columns: Worksheet columns

                 cure  nocure  All

ceftaroline       387      72  459
                372.05   86.95

ceftriaxone       349     100  449
                363.95   85.05

All               736     172  908

Cell Contents:      Count
                    Expected count


Pearson Chi-Square = 6.411, DF = 1, P-Value = 0.011
Likelihood Ratio Chi-Square = 6.431, DF = 1, P-Value = 0.011
```

This was obtained by making sure `Summarized data in a two-way table` was selected from the top drop-down list in the **Chi-Square Test for Association** dialogue box (**Stat** > **Tables** > **Chi-Square Test for Associaton**), entering `cure` and `nocure` into the **Columns containing the table** field and entering `treatment` into the **Rows** field.

The p-value for the test is 0.011, so by Table 1 (Subsection 6.2) there is moderate evidence that the treatment and being cured are not independent. Comparing the Expected values with those observed, more patients in the ceftaroline group were cured than expected, and fewer patients in the ceftriaxone were cured than expected. So there is a difference in pneumonia cure rate between the two treatment groups: ceftaroline appears to be more effective than ceftriaxone in curing pneumonia.

(c) This time the output produced by Minitab is as follows.

```
Rows: treatment   Columns: Worksheet columns

               diarrhoea  nodiarrhoea   All

ceftaroline           26          587   613
                   20.97       592.03

ceftriaxone           16          599   615
                   21.03       593.97

All                   42         1186  1228

Cell Contents:      Count
                    Expected count


Pearson Chi-Square = 2.499, DF = 1, P-Value = 0.114
Likelihood Ratio Chi-Square = 2.522, DF = 1, P-Value = 0.112
```

This is obtained in the same way as the output in part (a) was. The only difference is that this time `diarrhoea` and `nodiarrhoea` are entered in the **Columns containing the table** field.

This time there is little evidence against the null hypothesis ($p = 0.114$). So there is little evidence for a difference in the chances of diarrhoea between patients taking ceftaroline and ceftriaxone.

Solution to Computer activity 97

(a) This is a group-comparative trial. Each subject received either the standard treatment or the standard treatment plus stretching exercises, so it was not a crossover trial. Also, subjects in the trial were not specifically matched with other subjects in the trial, so it was not a matched-pairs trial.

(b) The 'changes in lateral rotation' are interval scale data as they correspond to measurements.

(c) Interval scale data from group-comparative trials can be analysed using two-sample t-tests.

(d) The hypotheses correspond to a two-sided test because the alternative hypothesis includes a 'not equals'. (For a one-sided test, the alternative hypothesis would have included either a 'greater than' or a 'less than'.)

(e) The output produced by Minitab is as follows.

```
Two-sample T for treatment vs control

            N  Mean  StDev  SE Mean
treatment  27  7.56   5.55      1.1
control    12  2.17   5.83      1.7


Difference = μ (treatment) - μ (control)
Estimate for difference:  5.39
95% CI for difference:  (1.43, 9.35)
T-Test of difference = 0 (vs ≠): T-Value = 2.76  P-Value = 0.009  DF = 37
Both use Pooled StDev = 5.6337
```

This output was obtained by selecting `Each sample is in its own column` from the top drop-down list of the **Two-sample t for the Mean** dialogue box (**Stat** > **Basic Statistics** > **2-Sample t**) and entering `treatment` and `control` in the **Sample 1** and **Sample 2** fields respectively. Also by making sure that the **Assume equal variances** option was selected in the **Two-Sample t: Options** dialogue box.

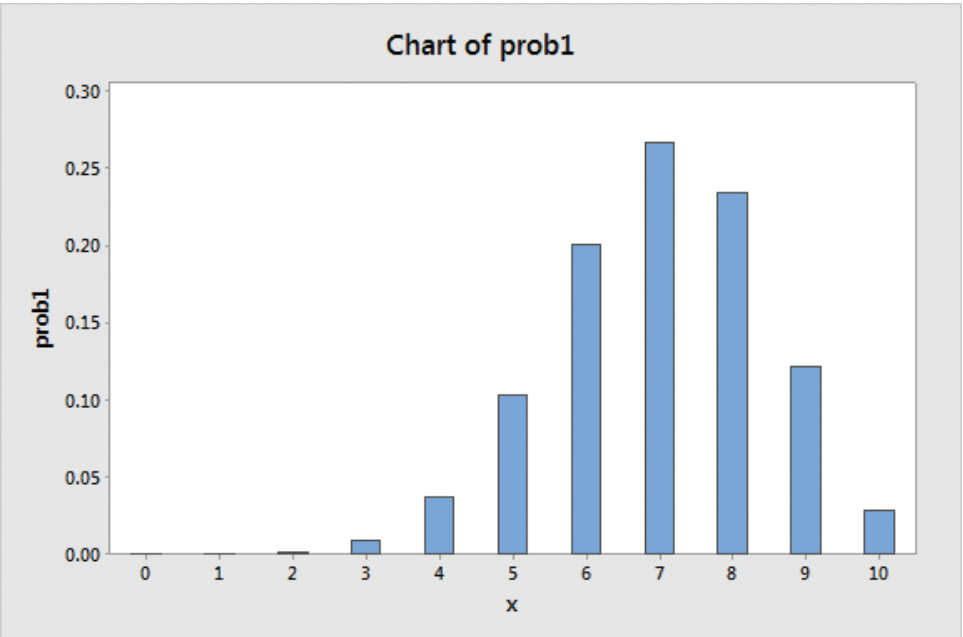
The p-value is between 0.01 and 0.001 so, using Table 1 (Subsection 6.2), there is strong evidence that the change in lateral rotation is not the same in the two groups. In fact, it appears that there is a greater change in lateral rotation when stretching exercises are given in addition to the standard treatment.

(f) From the output in part (e), the 95% confidence interval for $\mu_t - \mu_c$ is $(1.43, 9.35)$. This agrees with the conclusion drawn in part (e) as the interval is entirely above zero.

(g) From the output produced by Minitab, the standard deviations for the change in lateral rotation in the treatment and control groups are $s_t = 5.55$ and $s_c = 5.83$. Thus the ratio of the larger to the smaller variance is $5.83^2/5.55^2 \simeq 1.10$. As this value is less than 3, it is reasonable to have used a pooled standard deviation in parts (e) and (f).

Solution to Computer activity 99

(a) The bar chart of the probability distribution obtained in Computer activity 98 is as follows.



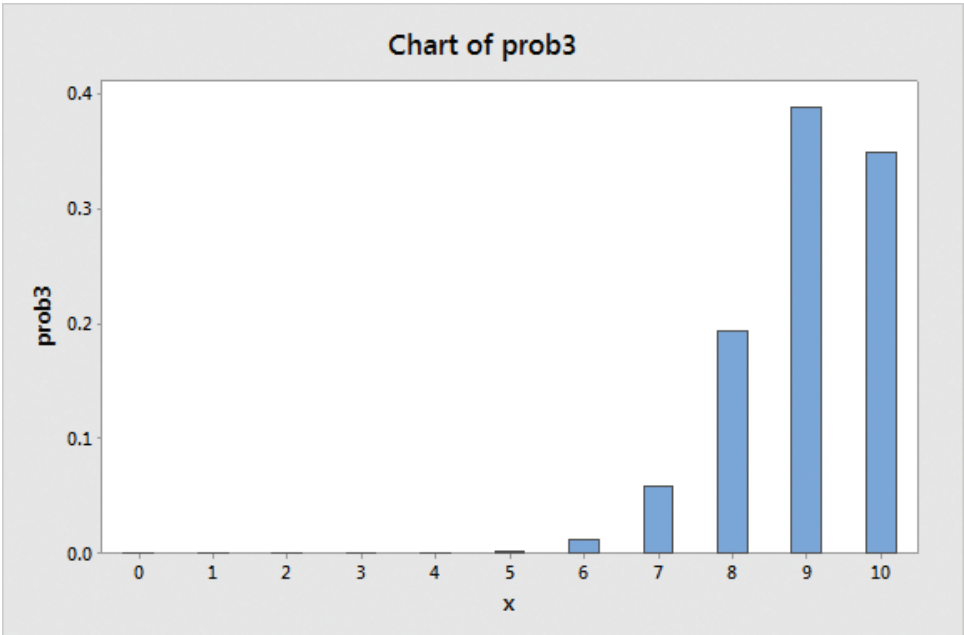
This bar chart was obtained by entering `prob1` in the **Graph variables** field and `x` in the **Categorical variable** field of the **Bar Chart: Values from a table, One column of values, Simple** dialogue box. (**Graph** > **Bar Chart**, select `Values from a table` for the **Bars represent** field and **Simple** as the form of bar chart.)

(b) The binomial distribution with $p = 0.7$ and $n = 10$ is unimodal, with the peak at about 7. In fact, all binomial distributions are unimodal and the peak is always near the value np. (In this case $np = 10 \times 0.7 = 7$.)

The binomial distribution with $p = 0.7$ and $n = 10$ is also left-skew.

Solution to Computer activity 100

(a) A bar chart of the probabilities is as follows.



The probabilities were obtained by entering the following in the **Binomial Distribution** dialogue box (**Calc** > **Probability Distributions** > **Binomial**, with **Probability** selected): `0.9` in the **Event probability** field, `10` in the **Number of trials** field, `prob3` in the **Optional storage** field and `x` in the **Input column** field. As an example, 0.057 396 is the probability that exactly 7 of the pupils pass in three or fewer attempts.

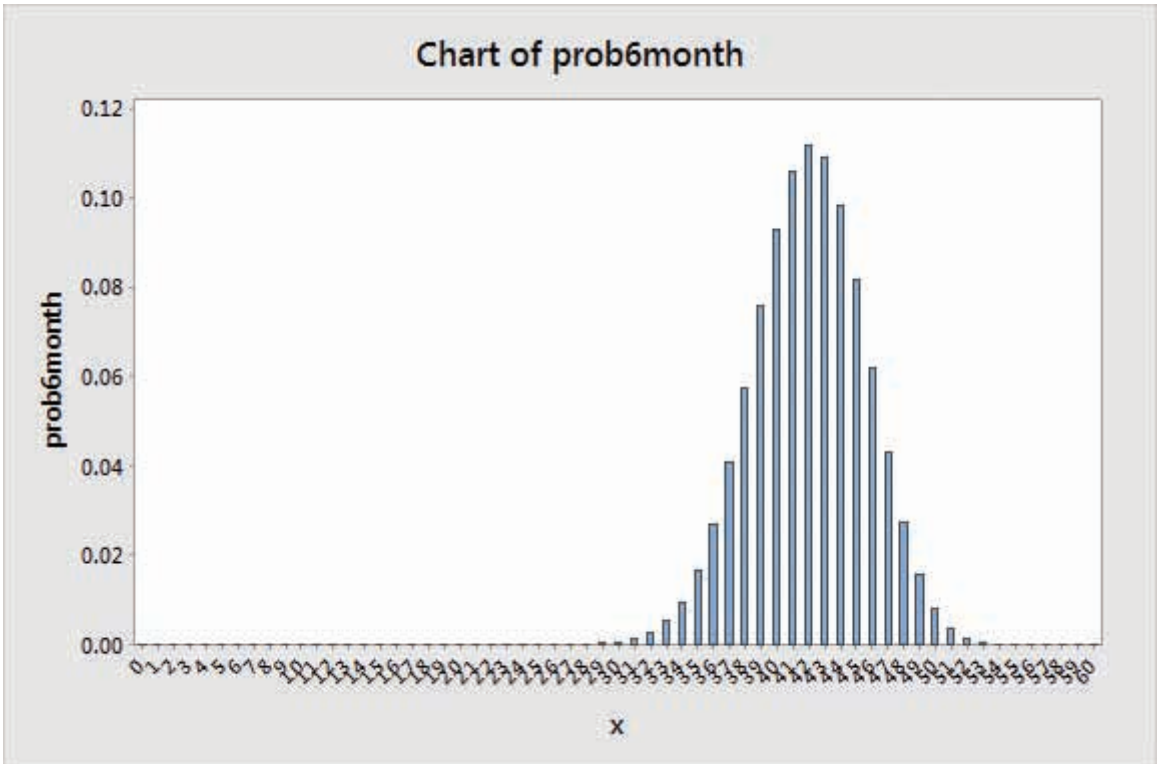
The bar chart was then obtained by entering `prob3` in the **Graph variables** field and `x` in the **Categorical variable** field of the **Bar Chart: Values from a table, One column of values, Simple** dialogue box. (**Graph** > **Bar Chart**, select `Values from a table` for the **Bars represent** field and **Simple** as the form of bar chart.)

(b) The binomial distribution with $p = 0.9$ and $n = 10$ is also unimodal, but this time the peak occurs around 9. This also corresponds to the value given by np.

The binomial distribution with $p = 0.9$ and $n = 10$ is more (left-)skew than the binomial distribution with $p = 0.7$ and $n = 10$.

Solution to Computer activity 101

(a) The bar chart that you should obtain is as follows.



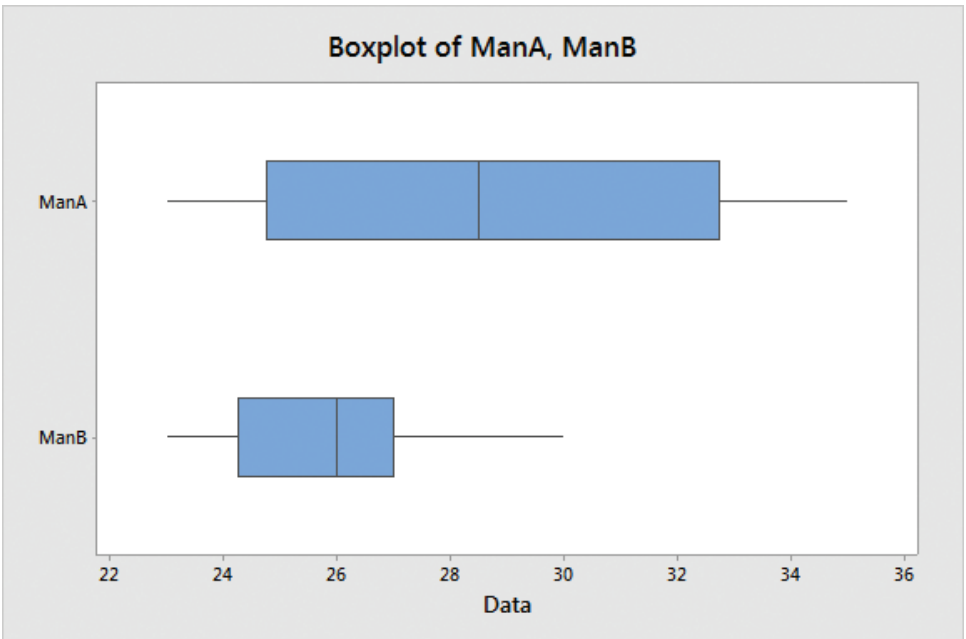
The probabilities of the binomial distribution with $p = 0.7$ and $n = 60$ were obtained following the same steps as in Computer activity 98, with the following exceptions: the values in the worksheet column labelled `x` run from `0` to `60`, the value `60` is entered in the **Number of Trials** field, and `prob6month` is entered in the **Optional storage** field. As an example, the 50th element of `prob6month`, corresponding to a value of $x = 49$, is 0.015 597.

The bar chart was then obtained following the same steps given in the solution to Computer activity 99, with the exception that `prob6month` was entered in the **Graph variables** field.

(b) The bar chart obtained in part (a) shows that the distribution is almost symmetric. For (i), note that it is much more symmetric than the binomial distribution with $p = 0.7$ and $n = 10$ (see solution to Computer activity 99). For (ii), note that the distribution looks bell-shaped, like a normal distribution.

Solution to Computer activity 102

(a) The diagram that you should obtain is as follows.



This diagram was obtained by entering both `ManA` and `ManB` in the **Graph variables** field in the **Boxplot: Multiple Y's, Simple** dialogue box (**Graph** > **Boxplot** and selecting **Multiple Y's, Simple**). Also, make sure that the **Transpose value and category scales** option is selected in the **Boxplot: Scale** dialogue box (obtained by clicking on the **Scale...** button).

(b)
- Neither boxplot has any extreme values and each sample has a distribution that seems reasonably symmetric. Thus, from the boxplots, the samples may come from populations that have normal distributions.
- The 'box' for the release times of tablets made by manufacturer A (`ManA`) is much bigger than the 'box' for the release times of tablets made by manufacturer B (`ManB`), and its whiskers cover a much bigger range (despite there being only 10 tablets in the sample from manufacturer A, while the sample from manufacturer B relates to 20 tablets). Hence the boxplots suggest it is unreasonable to believe that the samples come from populations whose variances are equal.

(c) The variance of `ManA` is 19.38 and the variance of `ManB` is 3.684.

These values were obtained using the **Display Descriptive Statistics** dialogue box (**Stat** > **Basic Statistics** > **Display Descriptive Statistics**), with the variables `ManA` and `ManB` added to the **Variables** field and making sure **Variance** is selected in the **Display Descriptive Statistics: Statistics** dialogue box (obtained by clicking on the **Statistics...** button in the **Display Descriptive Statistics** dialogue box).

The ratio of the bigger variance to the smaller variance is

$$\frac{19.38}{3.684} \simeq 5.26.$$

Now, the rule of thumb first given in Subsection 3.3 of Unit 10 says that the population variances can be treated as equal if the larger variance is less than three times the size of the smaller variance. That is not the case with `ManA` and `ManB`.

So it is not reasonable to assume that the samples come from populations whose variances are equal. The release times of tablets from manufacturer A appear to be more spread out than the release times of tablets from manufacturer B.

Solution to Computer activity 103

(a) The output from the two-sample t-test for populations with unequal variances is as follows.

```
Two-sample T for ManA vs ManB

        N   Mean  StDev  SE Mean
ManA   10  28.60   4.40      1.4
ManB   20  26.00   1.92     0.43


Difference = μ (ManA) - μ (ManB)
Estimate for difference:  2.60
95% CI for difference:  (-0.65, 5.85)
T-Test of difference = 0 (vs ≠): T-Value = 1.78  P-Value = 0.105  DF = 10
```

The test statistic, t, is given as 1.78. The value for the degrees of freedom is given as 10. Note that Minitab has used a complicated expression to come up with this value for the degrees of freedom, with truncation used to ensure that the final value is a whole number.

(b) Using the output given in part (a), the p-value is 0.105. According to Table 1 (Subsection 6.2), this corresponds to little evidence against the null hypothesis. In other words, there is little evidence that the tablets from the two manufacturers differ in the average time taken to release 50% of the active ingredient.

(c) Using the output given in part (a), the 95% confidence interval is $(-0.65\,\text{s}, 5.85\,\text{s})$. This interval contains the value 0, so the hypothesis test does not reject H_0 at the 5% significance level – consistent with the result found in part (b).

Solution to Computer activity 104

(a) The output from the two-sample t-test when the assumption of equal population variances is made is as follows.

```
Two-sample T for ManA vs ManB

       N   Mean  StDev  SE Mean
ManA  10  28.60   4.40      1.4
ManB  20  26.00   1.92     0.43


Difference = μ (ManA) - μ (ManB)
Estimate for difference:  2.60
95% CI for difference:  (0.26, 4.94)
T-Test of difference = 0 (vs ≠): T-Value = 2.27  P-Value = 0.031  DF = 28
Both use Pooled StDev = 2.9544
```

This output was obtained in the same way as in Computer activity 103, with one crucial difference: in the **Two-Sample t: Options** dialogue box, **Assume equal variances** was selected.

The value of the test statistic is 2.27. This is higher than the value obtained when equal variances were not assumed. This difference is because the value of the ESE assuming a common variance must be smaller than the value of the ESE not assuming a common variance.

The value of the degrees of freedom is given as 28. This corresponds, as it should, to $n_1 + n_2 - 2$, where n_1 and n_2 are the two sample sizes.

So in this case, the degrees of freedom used when the assumption of a common variance is made is much larger than when the assumption is *not* made. (You saw in Computer activity 103 that the degrees of freedom was only 10 when the assumption of a common variance was not made.)

(b) Using the output given in part (a), the p-value from the test is 0.031. Thus this test rejects H_0 at the 5% significance level (and at the 3.1% significance level). If the assumptions underlying this test were correct, then from Table 1 (Subsection 6.2) there would be moderate evidence against H_0 – that is, moderate evidence that the tablets from the two manufacturers differ in the average time taken to release 50% of the active ingredient.

(c) Using the output given in part (a), the results of the two hypothesis tests are substantially different. With this dataset, the correct test (not assuming population variances are equal) finds little evidence of a difference, while moderate evidence of a difference is found when the dubious (in this case) assumption of equal population variances is made.

Solution to Computer activity 105

(a) The following is the output from the test that does not assume equal population variances.

```
Sample   N  Mean  StDev  SE Mean
1       12  6.20   1.50     0.43
2       20  8.10   2.80     0.63


Difference = μ (1) - μ (2)
Estimate for difference:  -1.900
95% CI for difference:  (-3.457, -0.343)
T-Test of difference = 0 (vs ≠): T-Value = -2.50  P-Value = 0.018  DF = 29
```

The p-value from the test is 0.018. According to Table 1 (Subsection 6.2), this corresponds to moderate evidence against the null hypothesis. If the assumptions underlying the test are correct, there is moderate evidence that the population means differ. (The test assumes that observations come from populations that are approximately normally distributed.)

(b) The following is the output from the test that assumes the population variances are equal.

```
Sample   N  Mean  StDev  SE Mean
1       12  6.20   1.50     0.43
2       20  8.10   2.80     0.63


Difference = μ (1) - μ (2)
Estimate for difference:  -1.900
95% CI for difference:  (-3.694, -0.106)
T-Test of difference = 0 (vs ≠): T-Value = -2.16  P-Value = 0.039  DF = 30
Both use Pooled StDev = 2.4063
```

The p-value from this test is 0.039. According to Table 1 (Subsection 6.2), this corresponds to moderate evidence against the null hypothesis. If its underlying assumptions are correct, there is again moderate evidence that the population means differ. (This test makes the assumptions made in part (a), and additionally assumes that the population variances are equal.)

(c) Although the values of the test statistic for these samples are different (-2.50 and -2.16), the values of the degrees of freedom are very similar (29 and 30).

The two tests give p-values that are fairly different (0.018 and 0.039). However, the resulting conclusions are the same.

Notice that the smaller p-value is given by the test that does not assume the population variances are equal. (It was the other way round in Computer activities 103 and 104.)

(d) The two variances are $1.5^2 = 2.25$ and $2.8^2 = 7.84$. As

$$\frac{7.84}{2.25} \simeq 3.48 > 3,$$

it should not be assumed that the two population variances are equal. Thus the test used in part (a) (that is, not assuming equal population variances) is the test that should be used.

However, the choice of test is not critical. Either version leads to the same conclusion – moderate evidence that the two treatments A and B are not equally effective. Treatment A appears to be more effective as it has the shorter average time with symptoms.

Cover image: minxlj/www.flickr.com